Universidad de Valladolid

Escuela de Ingeniería Informática de Valladolid

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática

Mención Tecnologías de la Información

Análisis de la influencia de características vocales en modelos de detección de voces clonadas mediante IA

Autor:

D. Fernando Mateos Martínez

Tutor:

D. Blas Torregrosa García

Índice general

Resumen **1**

Abstract **2**

1. Introducción **3**

1.1. Contexto y justificación 3

1.1.1. Orígenes históricos y avances tecnológicos 3

1.1.2. Implicaciones sociales y éticas 4

1.1.3. Justificación de la importancia 5

1.2. Problema y relevancia del estudio 5

1.3. Objetivos 6

1.3.1. Objetivo principal: Identificación de características vocales relevantes para la detección de voces clonadas mediante análisis estadístico 6

1.3.2. Otros Objetivos 6

2. Marco teórico **8**

2.1. Introducción a la clonación de voces mediante IA 8

2.1.1. Definición y características de la clonación de voces 8

2.1.2. Historia y evolución de las tecnologías de síntesis de voz 8

2.2. Fundamentos de la clonación de voces 10

2.2.1. Síntesis de voz tradicional 10

2.2.2. Text-to-Speech (TTS) 13

2.3. Redes Neuronales y Aprendizaje Profundo 14

2.3.1. Arquitectura de Redes Neuronales 14

2.3.2. Neuronas Artificiales 14

2.3.3. Capas de la Red Neuronal 14

2.3.4. Pesos y Sesgos 15

2.3.5. Funciones de Activación 16

2.3.6. Capacidad de Aprendizaje y Generalización 17

2.3.7. Redes Neuronales Convolucionales (CNN) 17

2.3.8. Redes Neuronales Profundas (DNN) 18

2.3.9. Redes Neuronales Recursivas (RNN) 19

3. Procesamiento de audio **20**

3.1. Procesamiento de Señales de Audio 20

3.1.1. Digitalización del Audio 20

3.1.2. Preprocesamiento de Audio 21

3.1.3. Transformación de Señales: Espectrogramas y MFCCs 23

4. Parámetros de la voz **25**

4.1. Introducción a los Parámetros de la Voz 25

4.2. Métricas de rendimiento utilizadas 31

4.2.1. Precisión (Accuracy) 31

4.2.2. Sensibilidad (Recall) 31

4.2.3. Especificidad (Specificity) 31

4.2.4. Precisión Positiva 32

4.2.5. F1-Score[41] 32

4.2.6. AUC-ROC (Área bajo la curva ROC)[42] 32

4.3. Métricas de eficiencia y rendimiento computacional 32

4.3.1. Tiempo de Respuesta 32

5. Características y proceso de entrenamiento **34**

5.1. Conjunto de datos utilizados 34

5.1.1. Preprocesamiento y normalización de datos 34

5.1.2. División en conjuntos de entrenamiento, validación y prueba 34

5.2. Características Acústicas Utilizadas 35

5.2.1. MFCCs (Coeficientes Cepstrales en la Escala Mel) 35

5.2.2. Delta-MFCCs 35

5.2.3. Energía (RMS) 36

5.2.4. Relación Armónico-Ruido (HNR) 36

5.2.5. Formantes 37

5.2.6. Contraste Espectral[49] 37

5.2.7. Ancho de Banda Espectral 37

5.2.8. Tasa de Cruces por Cero 38

5.2.9. Combinaciones Estratégicas 38

5.3. Diseño Arquitectónico de la Red Neuronal 40

5.3.1. Estructura General del Modelo[50] 40

5.3.2. Funcionamiento Adaptativo 40

5.3.3. Configuración 41

5.3.4. Parámetros Ajustables 41

5.3.5. Ventajas de la Arquitectura 42

5.4. Proceso de Entrenamiento . . . 42
5.4.1. Carga y Procesamiento de Datos . . . 42
5.4.2. Normalización y División de Datos . . . 43
5.4.3. Construcción del Modelo . . . 43
5.4.4. Parada Temprana . . . 43
5.4.5. Entrenamiento del Modelo . . . 43
5.4.6. Métricas de Evaluación . . . 44

6. Análisis de Resultados **45**
6.1. Marco Experimental . . . 45
6.1.1. Métricas de Evaluación . . . 45
6.1.2. Características Acústicas Analizadas . . . 45
6.1.3. Configuraciones Arquitecturales . . . 46
6.2. Metodología de Análisis Estadístico . . . 47
6.2.1. Objetivo del Análisis Comparativo . . . 47
6.2.2. Protocolo de Validación Estadística . . . 47
6.2.3. Índice de Dominancia . . . 48
6.3. Análisis de Resultados por Métricas de Rendimiento . . . 49
6.3.1. Precision . . . 49
6.3.2. Recall . . . 52
6.3.3. Specificity . . . 55
6.3.4. Accuracy . . . 58
6.3.5. F1-Score . . . 61
6.3.6. AUC-ROC . . . 64
6.4. Conclusiones . . . 66

7. Mejoras, Limitaciones y Consideraciones Éticas **67**
7.1. Optimización . . . 67
7.1.1. Ampliación de Fuentes de Datos . . . 67
7.1.2. Variantes en la arquitectura . . . 67
7.2. Limitaciones . . . 68
7.2.1. Avance de la Tecnología . . . 68
7.2.2. Errores . . . 69
7.3. Ética del experimento . . . 69

Bibliografía **70**

Lista de Figuras **77**

Resumen

El aumento en la generalización del uso de la Inteligencia Artificial plantean un desafío como es la detección de voces clonadas creadas mediante técnicas de IA. Con el progreso en la tecnología de síntesis vocal, la clonación de voces se ha vuelto una amenaza importante al posibilitar la creación de imitaciones creíbles y sólidas de la voz de un individuo para propósitos potencialmente dañinos.

El propósito principal de este proyecto es investigar sobre cómo afectan a la eficacia de un modelo de detección las características vocales o conjunto de estas con el que se le entrene.

Este proyecto puede ser un avance para la generalización en el área de la seguridad y autenticación en aplicaciones de voz ofreciendo la mayor eficiencia para identificar y prevenir usos indebidos de la tecnología de síntesis de voz.

Abstract

The widespread adoption of Artificial Intelligence presents new challenges, such as the detection of cloned voices generated using AI techniques. With advancements in voice synthesis technology, voice cloning has become a significant threat, enabling the creation of convincing and robust imitations of an individual's voice for potentially harmful purposes.

The main objective of this project is to investigate how the vocal characteristics, or combinations thereof, used to train a detection model affect its effectiveness.

This project represents a step forward in the generalization of security and authentication in voice-based applications, offering improved efficiency in identifying and preventing the misuse of voice synthesis technology.

Capítulo 1

Introducción

1.1. Contexto y justificación

1.1.1. Orígenes históricos y avances tecnológicos

La historia de la clonación de voces tiene sus raíces en los primeros experimentos de síntesis de voz realizados a principios del siglo XX. Uno de los hitos iniciales fue el Voder, un dispositivo desarrollado en 1939 por los laboratorios Bell Labs. Este innovador aparato, operado mediante

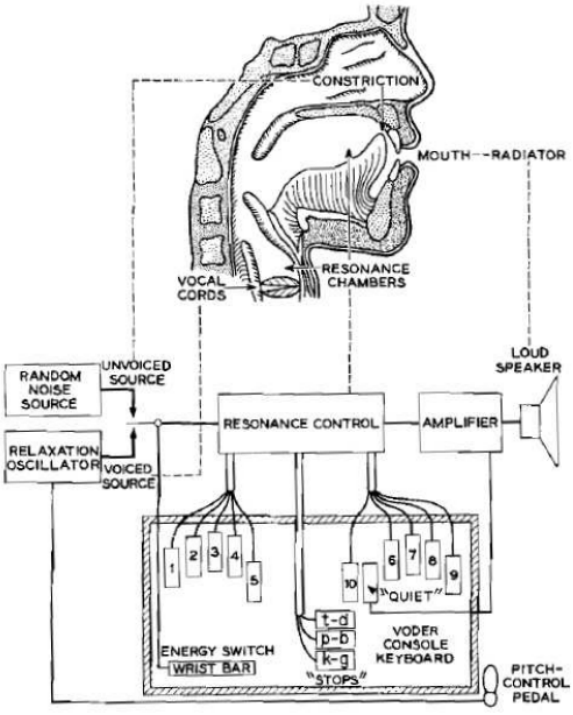


Figura 1.1: Diagrama Esquemático del sintetizador Voder obtenida de [1]

un sistema de teclas y pedales, logró imitar la voz humana generando sonidos básicos de manera mecánica. Aunque sus capacidades eran limitadas y requería gran habilidad para ser utilizado, el Voder supuso un avance crucial al demostrar que era posible recrear artificialmente la voz humana.

Estos primeros desarrollos no eran capaces de reproducir la complejidad y fluidez propias de la voz humana, pero sentaron las bases para la evolución de tecnologías más avanzadas. Poco después, surgió el Vocoder, un dispositivo que perfeccionó el procesamiento de señales de voz, mejorando la calidad y ampliando las aplicaciones de la síntesis vocal[1].

En las últimas décadas, los avances en aprendizaje automático y redes neuronales han revolucionado este campo. Un cambio radical llegó con la introducción de WaveNet, una red neuronal generativa desarrollada por DeepMind de Google en la década de 2010. WaveNet emplea redes neuronales convolucionales para generar voces de calidad notablemente similar a la humana. Este salto cualitativo fue posible gracias al incremento exponencial en la capacidad de procesamiento de datos y a la disponibilidad de grandes conjuntos de datos para entrenamiento.

Más adelante, en 2017, Google presentó Tacotron 2[2], un modelo que integraba WaveNet[3] con una red de secuencia a secuencia. Este sistema innovador convertía texto en espectrogramas de frecuencia, que luego eran transformados en audio mediante WaveNet. Esta combinación permitió una sincronización más precisa entre el texto y la voz generada, logrando resultados que se asemejan aún más a la voz humana real. Gracias a estos avances, la síntesis de voz ha alcanzado un nivel de realismo que parecía inalcanzable hace solo unas décadas.

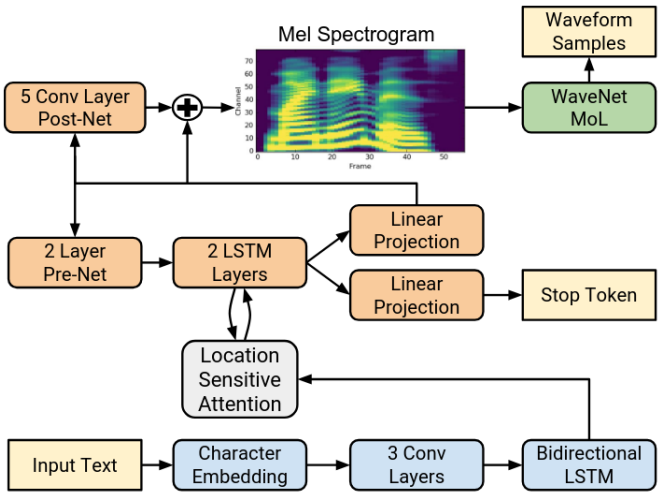


Figura 1.2: Diagrama de arquitectura del sintetizador TACOTRON2 de [2]

1.1.2. Implicaciones sociales y éticas

La clonación de voces con tanta precisión tiene implicaciones sociales y éticas bastante significativas, pues puede ser perjudicial para la privacidad y la seguridad [4]. La capacidad de replicar la voz de una persona permite, por ejemplo, la creación de *deepfakes* de audio que podrían usarse para engañar, cometer fraudes o difundir desinformación. Estos peligros no solo afectan a las personas, sino que también pueden comprometer a instituciones, medios de comunicación y sistemas de seguridad[5].

Esto nos plantea importantes preguntas sobre el consentimiento y los derechos. Es posible violar

los derechos a la privacidad y a la propiedad de la voz de una persona al crear una voz clonada sin su permiso. Además, el uso de voces clonadas para manipular opiniones o generar contenido falso plantea dilemas morales sobre la veracidad de la información y la confianza en las comunicaciones digitales.

1.1.3. Justificación de la importancia

El estudio y la detección de la clonación de voces son cruciales en el contexto actual por varias razones. Los sistemas de autenticación basados en la voz enfrentan importantes desafíos tecnológicos debido a los avances en la creación de voces clonadas. Es fundamental desarrollar técnicas de detección que evolucionen junto con los avances en síntesis de voz para proteger la integridad de los sistemas de seguridad y prevenir el abuso tecnológico.

Las voces clonadas pueden afectar la privacidad y la seguridad. La capacidad de replicar voces con alta fidelidad permite suplantar identidades, cometer fraudes y difundir desinformación. Estos usos muestran la necesidad de herramientas efectivas para detectar y reducir los efectos de la clonación de voces en la vida diaria. Comprender las técnicas de clonación y detección es fundamental para proteger la privacidad personal y la seguridad de las instituciones, así como para tomar las medidas preventivas adecuadas, cómo por ejemplo marcas de agua[6].

Por lo tanto, la investigación sobre la clonación y detección de voces a través de la IA es crucial tanto desde un punto de vista tecnológico como para abordar los problemas éticos y sociales que están surgiendo actualmente.

1.2. Problema y relevancia del estudio

La clonación de voces nos plantea un desafío creciente en campos desde el de la seguridad informática hasta el de la integridad de los datos que podemos encontrar en internet. El principal problema está en la capacidad de los sistemas más modernos que pueden replicar con alta fidelidad la voz humana, lo que puede ser utilizado de forma malintencionada para suplantar identidades[7], cometer fraudes o realizar ataques informáticos. A medida que las técnicas de síntesis de voz se han ido perfeccionando, es cada vez más complicado distinguir entre una voz real y una clonada, pudiendo explotar vulnerabilidades en los sistemas de autenticación que utilizan la voz como medio de verificación o comunicación.

Otro desafío es el de la suplantación de identidad; en los sistemas de autenticación basados en la voz, sobretodo cuando son usados en sectores críticos como el financiero o en dispositivos inteligentes, con las técnicas actuales de clonación de voces, se puede generar una réplica de la voz de una persona, permitiendo acceder a cuentas bancarias, realizar transacciones fraudulentas o manipular sistemas. Este tipo de ataques, conocidos como "deepfakes"de voz[8], son cada vez más frecuentes y más fáciles de realizar por gente con un mínimo conocimiento en el tema y esto si que es un problema.

No solo en deepfakes si no que también se pueden utilizar voces clonadas para engañar a empleados y hacerles creer que están recibiendo instrucciones de jefes, lo que se conoce como ataques de ingeniería social. Un ejemplo muy sonado ocurrió en 2019, cuando los atacantes usaron una voz clonada para hacerse pasar por un ejecutivo de una empresa y solicitar una transferencia de 243,000 dólares, y ocurrió, lo que demuestra que estas técnicas pueden llegar a funcionar.

En el ámbito de la política y los medios de comunicación, la clonación de voces puede utilizarse para difundir desinformación. Las voces clonadas pueden ser usadas para crear grabaciones falsas que impliquen a gente en declaraciones o acciones que nunca ocurrieron. Esta manipulación de la realidad a través de voces falsas puede poner en riesgo la integridad de los medios o para desgastar la confianza en distintas instituciones o personalidades.

El campo dónde más preocupante me parece es el de la privacidad. A medida que los sistemas evolucionan y las voces clonadas se vuelven más convincentes, las personas pierden el control sobre uno de los aspectos de su identidad: su voz. La capacidad de replicar voces sin su consentimiento acaba con la privacidad, ya que las personas pueden verse involucradas en situaciones (incluso ilegales)[9]donde sus voces son utilizadas sin su aprobación.

Dada la creciente sofisticación de estas tecnologías, es muy importante que se desarrollen y adopten métodos efectivos para detectar y contrarrestar estas amenazas, por esto la relevancia de estudiar este problema es indiscutible en el contexto actual, donde las implicaciones no tienen la magnitud que merecen y en muchos casos no se abordan cómo deberían.

1.3. Objetivos

1.3.1. Objetivo principal: Identificación de características vocales relevantes para la detección de voces clonadas mediante análisis estadístico

El objetivo principal de este TFG es analizar comparativamente distintas características vocales para determinar cuáles presentan diferencias estadísticamente significativas entre voces humanas auténticas y voces clonadas generadas mediante inteligencia artificial. El estudio busca establecer un marco metodológico que permita identificar los mejores marcadores de la voz, sentando las bases científicas para el desarrollo de sistemas de detección lo más eficientes posible.

A través de la ejecución y el análisis estadístico, se pretende cuantificar la capacidad discriminatoria de cada característica vocal, evaluando tanto su relevancia individual como su comportamiento combinado con otras características.

1.3.2. Otros Objetivos

1. **Seleccionar y categorizar características vocales clave:** Identificar un conjunto representativo de características acústicas (MFCCs, formantes, HNR, etc.) y parametrizar su

capacidad para distinguir entre voces reales y sintéticas mediante pruebas de significancia estadística.

2. **Proponer estrategias de optimización basadas en características:** Desarrollar recomendaciones para el diseño de sistemas de detección que prioricen las características más relevantes, estableciendo equilibrios óptimos entre precisión discriminativa (cómo el F1-Score) y coste computacional.

Capítulo 2

Marco teórico

2.1. Introducción a la clonación de voces mediante IA

La clonación de voces utilizando inteligencia artificial (IA) destaca cómo una de las aplicaciones más innovadoras dentro del campo. La perfección de esta tecnología ha generado tanto interés por sus posibles aplicaciones, como preocupaciones por cómo de mal se puede usar.

En este capítulo, se explica cómo funciona esta tecnología y cuáles son sus diferencias con respecto a otras, además de la historia y los avances que han llevado a las tecnologías actuales.

2.1.1. Definición y características de la clonación de voces

La clonación de voces es el proceso mediante el cual se replican las características vocales de un individuo. A través de distintas técnicas, se puede generar una voz artificial que imite los patrones específicos que tendría la voz de una persona, haciendo que la voz clonada parezca real y sin notar apenas cambios de tono, entonación, timbre y el resto de características. La clonación de voces busca capturar los matices que asemejen la voz clonada a la original.

A diferencia de la síntesis de voz tradicional, cuyo objetivo es generar una voz que suene lo más natural posible pero que sea genérica, la clonación de voces pretende replicar voces específicas. Las tecnologías más modernas analizan grabaciones de voz para identificar los patrones únicos de cada voz, ya que estos elementos son clave para hacer que la voz clonada se perciba como auténtica y no artificial.

2.1.2. Historia y evolución de las tecnologías de síntesis de voz

Los primeros intentos de emular la voz humana comenzaron con los primeros dispositivos que trataban de replicar los sonidos de las vocales. Obviamente estaba limitada en su capacidad para generar sonidos complejos, fue lo que inició el desarrollo de la síntesis de voz.

Uno de los hitos más importantes de la primera etapa fue la máquina "parlante"de Wolfgang von Kempelen, un dispositivo mecánico creado en el siglo XVIII que reproducía sonidos del habla

humana, principalmente las vocales, utilizando fuelles y tubos. Aunque era muy básica, esta máquina ayudó a comprender cómo se podrían generar sonidos similares al habla humana utilizando la mecánica.

Más tarde, durante el siglo XIX, las tecnologías de síntesis de voz experimentaron cambios importantes con el desarrollo de Euphonia , una máquina más avanzada basada en la máquina

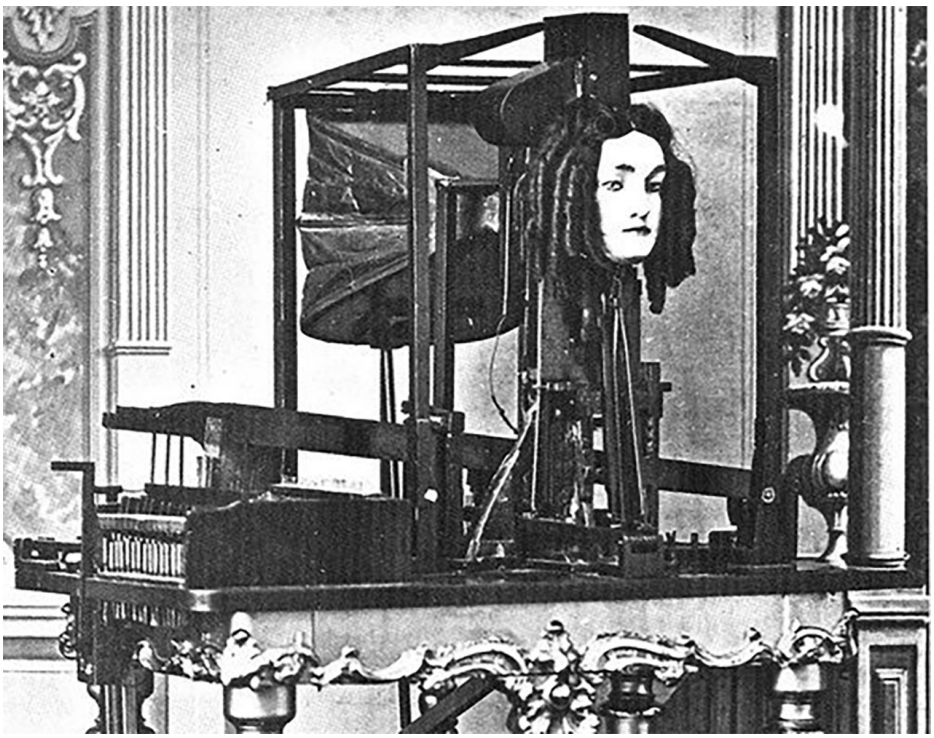


Figura 2.1: Fotografía de la máquina EUPHONIA[10]

parlante de Kempelen. Euphonia representó un avance significativo en la complejidad y precisión de la síntesis de voz, pues integraba conocimientos más profundos sobre el tracto vocal humano y su funcionamiento, aunque seguía teniendo limitaciones debido a dependía de la mecánica.

El verdadero gran salto se produjo en la década de los 30, con la invención del vocoder. Este dispositivo funcionaba como un analizador y sintetizador de voz y fue desarrollado en los laboratorios Bell. Aunque el habla que generaba el vocoder tenía una calidad robótica y nada de naturalidad, representó un hito en la reproducción de la voz pues utilizaba un análisis más detallado de las características del habla humana. A pesar de sus limitaciones en términos de realismo, sentó las bases para los avances posteriores.

Con el tiempo, el enfoque en la síntesis de voz cambió, pasando de utilizar dispositivos físicos a sistemas basados en modelos matemáticos y en lenguajes de programación. En las décadas de los 60 y 70, los modelos basados en **formantes** comenzaron a desarrollarse. Estos modelos se basaban en la idea de que la voz humana podía descomponerse en componentes fundamentales a los que se llama formantes, los cuales representaban picos de energía en el espectro de frecuencia de cada voz. Este enfoque basado en la informática permitió mejorar la precisión y naturalidad del habla sintetizada.

Con la llegada de la computación moderna y el desarrollo de muchas tecnologías digitales a finales del siglo XX marcó el auge de los sistemas de síntesis de voz por concatenación de unidades, como el DECTalk, popularizado en los años 80 y que utilizó Stephen Hawking para comunicarse. Aunque estos sistemas representaron avances, la flexibilidad de la concatenación seguía siendo limitada.

El verdadero salto cualitativo llega con el uso de la inteligencia artificial y las **redes neuronales** profundas a partir de la década de 2010. Tecnologías como WaveNet, desarrollada por DeepMind en 2016, revolucionaron el campo al permitir la generación directa de audio utilizando modelos basados en IA. WaveNet no solo mejoró significativamente la naturalidad del habla sintetizada, sino que también ofreció una mayor capacidad para capturar los matices del habla humana, como la entonación.

Otro avance reciente y notable fue Tacotron 2[2], un sistema de conversión de texto a voz desarrollado por Google. Tacotron 2 utiliza redes neuronales profundas para convertir el texto que se introduce en espectrogramas, que luego son transformados en formas de onda por un modelo parecido a WaveNet. Este enfoque permitió una síntesis de voz mucho más realista de las que se conseguían hasta el momento, pero sobre todo más flexible, ya que la concatenación de unidades podía ser muy realista pero no tenía flexibilidad y requería de una gran cantidad de datos.

Más recientes son otros modelos, como los desarrollados por OpenAI, que son capaces de clonar una voz con apenas unos segundos de muestra.

2.2. Fundamentos de la clonación de voces

2.2.1. Síntesis de voz tradicional

La síntesis de voz tradicional abarca una serie de métodos desarrollados antes de las técnicas basadas en inteligencia artificial. Estos métodos se centraban en generar una voz artificial que, aunque limitada en muchos aspectos, permitía la creación de sonidos que eran comprensibles para los oyentes. Los principales enfoques son la síntesis de formantes y la síntesis por concatenación.

Síntesis de Formantes

Uno de los primeros métodos que consiguió buenos resultados para la generación de voz artificial fue la síntesis de formantes. Este enfoque se basa en la idea de que los sonidos del habla pueden descomponerse en componentes acústicos llamados formantes[11], que son los picos de energía en el espectro de frecuencias de la voz. Estos formantes se vuelven fundamentales para la percepción de las vocales y consonantes, ya que corresponden a las resonancias naturales del tracto vocal humano.

Características

La síntesis de formantes utiliza modelos acústicos que imitan la fisiología del tracto vocal humano. Los formantes se ajustan para reproducir las características de los diferentes sonidos, en

función de su frecuencia y amplitud. A diferencia de las grabaciones reales de voz, este método genera la señal de habla desde cero, simulando las cuerdas vocales y una serie de filtros que ayudan a modelar las resonancias del tracto vocal.

Un ejemplo notable de síntesis de formantes es el dispositivo "VODER" , del cual hemos hablado

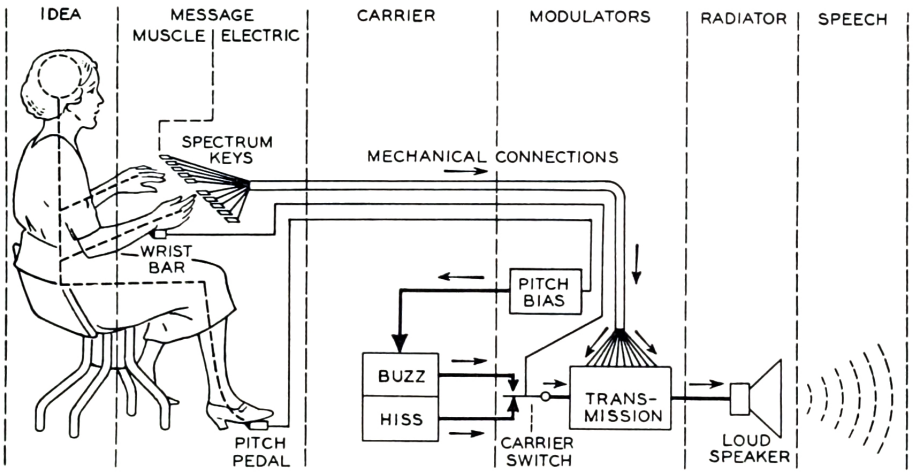


Fig. 8—Schematic circuit of the voder.

Figura 2.2: Circuito esquemático del Voder[12]

anteriormente pues fue capaz de generar sonidos de voz controlados manualmente mediante un teclado.

Limitaciones

A pesar de la innovación que suponía, la síntesis de formantes presentaba varias limitaciones:

- **Falta de naturalidad**: La voz generada solía sonar robótica y artificial, ya que no capturaba muchos de los matices de la voz humana, como la emoción o la variabilidad natural a lo largo de una frase.
- **Complejidad en el modelado**: Era necesario ajustar una gran cantidad de parámetros para obtener una síntesis de voz que al menos fuera aceptable, lo que hacía el proceso complejo y poco flexible.

A pesar de estas limitaciones, la síntesis de formantes sentó las bases *conceptuales* para avances posteriores al introducir la idea de la descomposición de la voz.

Síntesis por Concatenación de Unidades

La síntesis por concatenación de unidades representó un avance respecto a los métodos basados en formantes. En lugar de generar el habla desde cero, este enfoque utiliza fragmentos pregrabados de voz humana, los cuales se concatenan para formar las palabras y oraciones[13]. Estos fragmentos pueden ser fonemas individuales, sílabas, palabras o incluso frases completas, dependiendo del sistema.

Características

- **Uso de grabaciones reales**: A diferencia de la síntesis de formantes, este método se basa en grabaciones reales de voz humana, lo que proporciona una mayor naturalidad.

- **Segmentación y selección de unidades**: Uno de los principales desafíos es encontrar la manera más óptima para segmentar las grabaciones en unidades manejables y seleccionar las más apropiadas en cada contexto. Se utilizan técnicas de *alineación fonética* para lograr una concatenación fluida entre los fragmentos.

- **Variaciones en el tamaño de las unidades**: Existen variantes en el tamaño de las unidades de habla utilizadas. Por ejemplo, en la síntesis por concatenación *diphone*, se emplean secuencias de dos fonemas consecutivos, lo que mejora la fluidez en la transición entre sonidos.

Aplicaciones y Ejemplos

Uno de los sistemas más conocidos basados en la síntesis por concatenación fue *DECTalk*, que aunque no era perfecto, ofrecía una mayor naturalidad que los enfoques basados en formantes.

Limitaciones

A pesar de los avances en calidad, la síntesis por concatenación enfrentaba varias limitaciones:

- **Dependencia de las grabaciones**: La calidad de la voz depende en gran medida de la diversidad de las grabaciones que se tengan en las bases de datos. Esto hacía que las voces creadas fueran menos flexibles, ya que había muchas dificultades para generar pronunciaciones o expresiones no bien representadas.

- **Problemas de fluidez**: Aunque la calidad de los fragmentos era alta, la concatenación de unidades a menudo resultaba en transiciones bruscas o artificiales entre los fragmentos.

- **Tamaño de almacenamiento**: La necesidad de almacenar grandes bases de datos de grabaciones suponía un desafío en cuanto a los requisitos de memoria, especialmente cuando el almacenamiento era más limitado que en la actualidad.

Impacto y Legado

La síntesis de formantes y la síntesis por concatenación de unidades han sido tecnologías clave para el desarrollo de la síntesis de voz. Aunque estos métodos han quedado en gran parte obsoletos con la llegada de las técnicas basadas en inteligencia artificial, como las redes neuronales profundas, la influencia ha sido muy grande y se utilizan en algunos contextos. Por ejemplo, algunos sistemas híbridos actuales combinan elementos de síntesis por concatenación con técnicas avanzadas para mejorar la eficiencia y calidad de la voz sintética.

2.2.2. Text-to-Speech (TTS)

La tecnología Text-to-Speech (TTS) tiene como objetivo convertir un texto de entrada en voz hablada. Para lograrlo, realiza una serie de transformaciones, comenzando desde el análisis del texto escrito hasta la síntesis de la señal de audio que imita la voz.

El proceso se divide generalmente en dos fases[14]:

Análisis del texto

En esta etapa, el sistema procesa el texto de entrada para identificarlo y convertirlo en la representación fonética. Esto implica convertir cada palabra en una secuencia de sonidos o fonemas, que son las unidades básicas del habla. Además, el sistema debe estar programado para distinguir correctamente la puntuación, acrónimos, fechas, números y otras peculiaridades que tiene el lenguaje escrito para poder crear la representación de una manera correcta y sin fallos.

Síntesis de voz

Una vez que el texto se ha convertido en una representación fonética, el sistema genera la señal de audio correspondiente a esos fonemas. El método de síntesis puede ser cualquiera de los que se han mencionado anteriormente o integrar técnicas basadas en inteligencia artificial, cómo veremos más adelante.

La calidad de un sistema TTS se mide en función de la claridad de la voz sintetizada y la fluidez con la que reproduce el texto de manera cercana a cómo lo haría una persona.

Evolución y mejoras en TTS

A lo largo del tiempo, los sistemas TTS han evolucionado para mejorar la naturalidad y comprensibilidad de las voces sintetizadas. Los primeros sistemas TTS eran útiles para tareas como la lectura de textos para personas con discapacidades, pero las voces resultaban bastante artificiales y planas y no eran lo esperado, por ejemplo, para una clonación de voz.

La introducción de mejores técnicas para capturar la prosodia ha sido un elemento clave para mejorar la calidad de los sistemas TTS. La prosodia se refiere a los cambios en el tono, la duración y el ritmo del habla, que tienen un impacto en cómo se recibe el significado independientemente de las palabras. El mejor control de la prosodia ha hecho que las voces suenen más naturales y menos monótonas[15].

Aplicaciones y avances recientes

El uso de sistemas TTS se ha expandido considerablemente en los últimos tiempos y ahora es una tecnología muy utilizada en varios contextos:

- **Asistentes virtuales**: Sistemas como Alexa (Amazon) y Siri (Apple) dependen de la tecnología TTS para interactuar con los usuarios.
- **Lectores de pantalla**: Herramientas para leer en voz alta documentos, correos electrónicos y páginas web, lo que mejora la accesibilidad de aplicaciones y sitios web.
- **Sistemas de respuesta automática**: Los sistemas de respuesta de voz interactiva utilizan TTS para proporcionar información y gestionar llamadas de clientes de manera eficiente.

2.3. Redes Neuronales y Aprendizaje Profundo

[16][17][18] Las redes neuronales han transformado profundamente el campo de la inteligencia artificial, incluida la clonación de voces[19]. A través de estas arquitecturas, los sistemas que utilizan la IA han mejorado muchísimo su capacidad para analizar y reproducir los patrones más complejos, como el del habla humana, con niveles de precisión que no se habían llegado a alcanzar con los métodos anteriores.

El uso de redes neuronales ha permitido que estos sistemas sean capaces de aprender de grandes volúmenes de datos de voz. Para entender cómo estos avances han sido posibles, es importante entender los diferentes tipos de redes neuronales, cada una con capacidades y enfoques específicos para procesar los datos y cómo funcionan por dentro.

2.3.1. Arquitectura de Redes Neuronales

[20]

La arquitectura de una red neuronal está inspirada en el funcionamiento del cerebro humano, imitando el comportamiento de las neuronas para procesar y transmitir la información. Las redes neuronales consisten en una serie de componentes interconectados que trabajan en conjunto y son capaces de aprender patrones complejos a partir de los datos que reciben, lo que las hace encajar en aplicaciones como el reconocimiento de voz o la clonación de voces. A continuación, se explican los principales componentes que podemos encontrar en un sistema con arquitectura de Redes Neuronales.

2.3.2. Neuronas Artificiales

La unidad fundamental de una red neuronal es la **neurona**, aunque también se le puede denominar **nodo**. Cada neurona recibe varias entradas o (*inputs*), las procesa y produce una salida. Esta salida se transfiere cómo una entrada a otras neuronas en capas posteriores, siguiendo un esquema de conexiones.

El nodo o neurona recibe entradas $x_1, x_2, \ldots, x_n$, cada una de ellas multiplicada por un peso $w_1, w_2, \ldots, w_n$. Los pesos son coeficientes ajustables que determinan la importancia de cada entrada. Además, cada neurona puede tener un **sesgo** (*bias*), que ajusta el resultado. El valor obtenido se pasa a través de una **función de activación**, que proporciona no linealidad, permitiendo que las redes modelen datos más complejos.

2.3.3. Capas de la Red Neuronal

Una red neuronal típica está organizada en **capas**, donde cada capa contiene un conjunto de neuronas. Hay tres tipos principales de capas:

Capa de entrada

La capa de entrada es la primera capa de una red neuronal y su función principal es recibir los datos para comenzar su procesamiento. Cada neurona de esta capa es una variable o característica del conjunto de datos. Por ejemplo, en un problema de reconocimiento de voz, las neuronas de la capa de entrada pueden representar las características acústicas que vamos a extraer de una señal de audio, como las frecuencias o la intensidad. La capa de entrada simplemente transmite la información de los datos a las capas posteriores para su procesamiento.

Capas ocultas

Las capas ocultas se encuentran entre la capa de entrada y la capa de salida. A través de la aplicación de funciones de activación y el ajuste de pesos, estas capas permiten que la red neuronal aprenda patrones complejos en los datos. Cada neurona en una capa oculta está conectada a las neuronas de la capa anterior, y su salida es una mezcla ponderada (usando los pesos) de las entradas que recibe. Una red neuronal puede modelar relaciones no lineales y representaciones más abstractas de los datos de entrada cuando agregamos múltiples capas ocultas. Esto le permite resolver problemas más complejos, como detectar características en imágenes o identificar patrones en el habla.

Capa de salida

La capa de salida es responsable de generar la predicción o el resultado final. El número de neuronas presentes en esta capa varía según el tipo de tarea que se esté realizando. Cada neurona en la capa de salida representa una categoría o clase en un problema de clasificación, y la salida de la red es una probabilidad asignada a cada clase. Por ejemplo, en tareas de regresión, la capa de salida produce un valor numérico. Esta capa traduce la información procesada por las capas ocultas en un formato que se entienda, para que el usuario o la aplicación puedan representar los resultados.

2.3.4. Pesos y Sesgos

Los **pesos** son uno de los componentes más importantes en una red neuronal. Determinan la importancia que va a tener cada entrada, permitiendo que la red ajuste su comportamiento durante el entrenamiento para optimizar sus predicciones. Durante el proceso de entrenamiento, los pesos van ajustando utilizando algoritmos de optimización, lo que permite a la red mejorar su rendimiento y minimizar el error de predicción.

Aunque nosotros creemos que asignamos los pesos de manera correcta será la red la que según aprende los modifique; por ejemplo, en las muestras de voz de una persona, si el timbre es constante pero el ritmo o velocidad a la que habla varía significativamente, los pesos asignados a las

características relacionadas con el ritmo irán cambiando para ser más altos, permitiendo que la red capture esta variabilidad y mejore.

Los **sesgos** (*biases*), por otro lado, son valores que se añaden y que permiten desplazar la función de activación, facilitando que la red se ajuste a patrones en los datos que no están centrados alrededor de un origen. Por ejemplo, si la voz de la persona tiene un timbre muy grave en todas las muestras, el sesgo podría ajustar la voz para generar ese timbre con mayor precisión, independientemente de la variación de otras características.

2.3.5. Funciones de Activación

Las funciones de activación son los componentes clave en las redes neuronales, ya que sin ellas, las redes neuronales estarían limitadas en su capacidad para modelar relaciones complejas. En el contexto tanto de la clonación de voces como de la detección de voces clonadas, las funciones de activación permiten que la red neuronal identifique y distinga características esenciales que no serían evidentes si solo pudieran ver una representación lineal del audio.

En una red neuronal, cada neurona realiza dos operaciones básicas: calcular una combinación ponderada de las entradas y luego aplicar una función de activación a esta combinación. El resultado de esta función de activación es lo que pasa a la siguiente capa de la red. Sin una función de activación, las neuronas simplemente sumarían los pesos y sesgos.

A continuación se describen algunos tipos de funciones de activación:

- **ReLU (Rectified Linear Unit)**: Una de las funciones de activación más utilizadas en redes neuronales profundas es ReLU, que convierte cualquier entrada negativa en cero, manteniendo los valores positivos tal como están. Es muy simple , lo que la hace fácil de usar para entrenar redes profundas en tareas de procesamiento de audio.

- **Sigmoide**: La función sigmoide convierte cualquier entrada en un valor entre 0 y 1, lo que es útil cuando se necesita una clasificación binaria. En el caso de la clonación y detección de voces, esta función se puede utilizar en las capas finales para determinar si una voz es real o clonada. La salida es un valor entre 0 y 1, que se puede interpretar como la probabilidad de que la voz sea clonada.

- **Tanh (Tangente hiperbólica)**: La función tanh es otra función de activación no lineal que transforma los valores de entrada a un rango entre -1 y 1. Es útil para modelos que requieren que los valores negativos también tengan un significado. Para la detección de voces clonadas, esta función puede ayudar a identificar patrones acústicos raros que no serían evidentes si solo se permitieran valores positivos, mejorando la precisión de detección y encontrando peculiaridades en una voz.

2.3.6. Capacidad de Aprendizaje y Generalización

La capacidad de una red neuronal para aprender depende de su capacidad para ajustar los pesos y sesgos a través de la **retropropagación**, un proceso iterativo que reduce el error de la predicción. Este procedimiento se utiliza en el entrenamiento de la red utilizando un gran conjunto de datos etiquetados. La red ajusta los pesos a través de ciclos de entrenamiento para reducir los errores en sus predicciones.

Una red neuronal bien entrenada no solo debe aprender patrones de datos específicos, sino que también debe ser capaz de hacer predicciones precisas sobre nuevos datos no utilizados durante el entrenamiento. Esta capacidad es esencial para garantizar que los modelos sean efectivos y se puedan aplicar al mundo real[21].

2.3.7. Redes Neuronales Convolucionales (CNN)

Las **Redes Neuronales Convolucionales (CNN)** son un tipo de red neuronal que ha demostrado ser eficaz para el procesamiento de datos como las imágenes o el audio. Mientras que las redes neuronales tradicionales tratan cada entrada como una entidad independiente, las CNN se centran en capturar **patrones espaciales** y **relaciones locales** dentro de los datos, lo que las convierte en una muy buena herramienta para detectar características en diferentes tipos de secuencias. El componente clave que diferencia a las CNN de otras arquitecturas es la **operación de convolución**, un proceso que permite a la red aprender características locales, como podrían ser los bordes en una imagen o lo que más nos interesa(frecuencias en una señal de audio), que luego pueden ser utilizadas para reconocer patrones más complejos.

Estructura de las Redes Neuronales Convolucionales

Las CNN están compuestas por varias capas organizadas jerárquicamente, y cada una de estas capas tiene un papel específico en el procesamiento de la información:

- **Capas de Convolución**: En una CNN, la primera capa (y por lo general las siguientes a esta) son capas de convolución. Estas capas aplican un conjunto de **filtros** sobre la entrada, con el objetivo de extraer características de los datos de entrada. Para un espectrograma de audio, por ejemplo, los filtros pueden detectar patrones relacionados con la frecuencia o la amplitud. Estos filtros se van ajustando automáticamente durante el proceso de entrenamiento, lo que permite a la red aprender qué características más relevantes para la tarea para la que se este entrenando.

- **Capas de Agrupación**: Para reducir la cantidad de datos que se deben procesar y concentrar la información que de verdad es relevante, las CNN utilizan capas de **pooling o agrupación**, que son responsables de reducir las dimensiones de las características extraídas, pero manteniendo al mismo tiempo las más importantes. Este proceso es esencial para

hacer que el modelo sea más eficiente y no se pierda en características muy poco importantes, ocupando recursos que no son necesarios.

- **Capas Completamente Conectadas**: Al igual que en las redes neuronales tradicionales, al final de una CNN suelen añadirse una o más capas completamente conectadas. Estas capas toman las características extraídas por las capas anteriores y transforman las características aprendidas a un formato que pueda usarse para tareas de clasificación o de regresión.

Ventajas de las CNN

Una de las **principales ventajas** de las CNN es su capacidad para **capturar patrones locales** en los datos, lo que es particularmente útil para tareas como el análisis de imágenes. Otra ventaja es la **reducción de parámetros** comparada con otras redes tradicionales, debido al uso de filtros en las capas de convolución. Esto hace que las CNN sean más eficientes en términos de memoria y cómputo, necesario para procesar grandes volúmenes de datos.

Desafíos de las CNN

Uno de los principales desafíos es la necesidad de grandes cantidades de **datos etiquetados** que son necesarios para entrenar modelos efectivos. Además, las CNN tienden a ser más sensibles al **preprocesamiento de los datos**, como la normalización y la transformación de las señales de audio. Aunque las CNN capturan patrones locales con éxito, pueden tener dificultades para modelar dependencias a largo plazo en los datos. Por ello, se combinan con frecuencia con otros tipos de redes, como las RNN, para poder capturar mejor las relaciones locales y globales en datos secuenciales.

2.3.8. Redes Neuronales Profundas (DNN)

Las redes neuronales profundas son un subtipo de redes neuronales artificiales, pero tienen una característica esencial que las distingue de las redes neuronales más básicas: su profundidad. Las DNN tienen múltiples capas ocultas, lo que les permite modelar relaciones más complejas. Esto contrasta con las redes neuronales tradicionales o superficiales, que solo tienen una o pocas capas ocultas y solo pueden modelar relaciones relativamente sencillas. Las DNN pueden capturar patrones o información de mayor abstracción en los datos gracias a esta arquitectura profunda.

Estructura y Funcionamiento de las Redes Neuronales Profundas

Una DNN sigue la misma estructura básica que una red neuronal artificial simple, pero con un mayor número de capas ocultas interconectadas. La cantidad de capas ocultas varía mucho según la complejidad de la tarea para la que este configurada la red. Cada capa de DNN captura representaciones de datos de entrada más abstractas, lo que permite aprender características jerárquicas complejas.

Profundidad de la Red y su Impacto en el Rendimiento

La profundidad de una red, es decir, el número de capas ocultas que contiene, influye directamente en la capacidad de esta para aprender patrones complejos. Sin embargo, este aumento en la profundidad también tiene varios problemas, como el **sobreajuste** y el **desvanecimiento del gradiente**. Las redes profundas tienen una gran capacidad para extraer patrones complejos de los datos de entrenamiento, pero pueden ajustarse demasiado bien a los datos de entrenamiento, lo que afecta el rendimiento con nuevos datos. Además, en redes muy profundas, el desvanecimiento del gradiente puede dificultar el aprendizaje de relaciones a lo largo de las múltiples capas.

2.3.9. Redes Neuronales Recursivas (RNN)

Las redes neuronales recursivas (RNN) son un tipo de red neuronal diseñada para procesar secuencias de datos o series temporales, lo que las hace muy útiles para tareas como el análisis de secuencias de texto. Las RNN tienen la capacidad de mantener una "memoria"de entradas anteriores, lo que las diferencia de las redes neuronales tradicionales, que procesan las entradas de forma aislada o local. Esto es posible gracias a su estructura interna, en la que las conexiones entre las neuronas permiten que las salidas de una capa se retroalimenten como entradas para la siguiente capa de la red, creando un ciclo.

Aplicaciones de las RNN

Por ejemplo, en la clonación de una voz, la secuencia de palabras y sonidos no puede ser tratada de manera aislada. El tono, la entonación y la pronunciación de cada palabra dependen de las palabras anteriores en una frase. Las RNN son capaces de **modelar estas relaciones temporales**, lo que las convierte en una gran herramienta para capturar las dinámicas más complejas de la voz.

Desafíos de las RNN

Una de las limitaciones de las RNN básicas es su dificultad para recordar información a lo largo de secuencias largas. Existen variantes más avanzadas, como las **Long Short-Term Memory (LSTM)[22]** o las **Gated Recurrent Units (GRU)**, que introducen mecanismos para que la red decida qué información debe recordarz qué debe .ºlvidar". Aun así, las RNN, LSTM y GRU presentan otros desafíos, como el **desvanecimiento del gradiente** y su **ineficiencia** para procesar secuencias largas o en tiempo real.

Capítulo 3

Procesamiento de audio

3.1. Procesamiento de Señales de Audio

El procesamiento de señales de audio es un componente esencial en la detección de voces clonadas. Gracias a ello vamos a poder transformar las señales de audio en datos que pueden ser analizados para extraer información. En el contexto de la detección de voces clonadas, el correcto procesamiento de señales juega un papel importante que puede marcar la diferencia entre revelar características que pudieran ser imperceptibles para el oído humano o que esto no ocurra. A través del análisis detallado de las señales de audio, será más fácil que las redes puedan capturar patrones y anomalías que podrían indicar una manipulación de la voz. Esto permite que los sistemas de IA también identifiquen irregularidades en la estructura o el comportamiento de la señal que lo haga parecer generado.

3.1.1. Digitalización del Audio

La **digitalización de audio** es el primer paso y uno de los más importantes ya que permite convertir las señales de audio analógicas (nuestro habla) en datos digitales que pueden ser procesados por los algoritmos de inteligencia artificial. Este proceso se realiza mediante el **muestreo** y la **cuantificación**, que transforman una señal continua en una serie de valores que los sistemas pueden analizar mejor para identificar si una voz ha sido manipulada o generada artificialmente[23].

Muestreo

El **muestreo** es el proceso mediante el cual se toman muestras de la señal de audio (que es analógica) en intervalos de tiempo regulares. La frecuencia con la que se toman estas muestras se denomina **tasa de muestreo** (sampling rate), y se mide en hercios (Hz). Para capturar y reconstruir correctamente una señal de audio sin que perdamos información la tasa de muestreo debe ser al menos el doble de la frecuencia más alta que existe en la señal original (según el

teorema de Nyquist[23]). Por ejemplo, si una señal de voz contiene frecuencias de hasta 10 kHz, la tasa de muestreo mínima necesaria sería de 20 kHz. No cumplir con esta condición puede llevar a la pérdida de información o a la aparición de distorsiones, lo que haría que los datos que se llevan a la red neuronal no fueran precisos.

Las voces generadas artificialmente pueden presentar irregularidades en las frecuencias más altas o más bajas de la señal, y por eso es importante evitar estas pérdidas. Estas anomalías podrían ser difíciles de detectar si la señal no se muestrea correctamente. Las tasas de muestreo más comunes en aplicaciones de audio son 44.1 kHz y 48 kHz, que son suficientes para capturar la mayoría de los detalles del habla humana. Sin embargo, en aplicaciones más avanzadas de detección de voces clonadas, podrían ser necesarias tasas más altas. Si eliminamos esas frecuencias estamos perdiendo capacidad de detección.

Cuantificación y Resolución de Bits

Una vez muestreada, la señal analógica debe ser **cuantificada**, es decir, convertida en valores numéricos que representen la amplitud de la señal en cada punto. La **resolución de bits** va a determinar cuántos valores diferentes puede tomar cada muestra, lo que influye directamente en la calidad del audio que digitalizamos. Una mayor resolución de bits permite representar los niveles de amplitud con mayor precisión[24], lo que es esencial para captar los matices más sutiles y las anomalías.

Una voz clonada puede parecer natural en un análisis superficial, pero anomalías como fluctuaciones "raras" en la amplitud o ruido inesperado no podrían ser detectadas si no contáramos con una resolución de bits adecuada. Una resolución baja podría ocultar estas alteraciones y hacer que la voz clonada pase desapercibida, mientras que una resolución alta permite al algoritmo de detección que se programe realizar análisis más detallados.

Impacto de la Tasa de Muestreo y la Resolución en la Detección

La **tasa de muestreo** y la **resolución de bits** afectan directamente a la eficacia de los algoritmos de detección de voces clonadas. Aunque una mayor resolución de bits y tasa de muestreo incrementan el coste computacional, es necesario que estos valores sean lo suficientemente altos para que el sistema funcione de manera óptima.

3.1.2. Preprocesamiento de Audio

El **preprocesamiento de audio** se encarga de preparar las señales para su análisis. Las señales de audio sin ningún procesamiento previo suelen contener ruido, fluctuaciones de volumen, pausas prolongadas u otras imperfecciones que pueden interferir con la capacidad del sistema para detectar con precisión si esa voz ha sido generada artificialmente o no. A través del preprocesamiento[25],

se aplican diversas técnicas para "limpiar"la señal, mejorando significativamente la precisión y reduciendo los fallos que pueda tener el sistema.

Normalización

La **normalización** es una técnica que se utiliza para ajustar el nivel de amplitud de una señal[26] de audio, con el objetivo de que todas las muestras de voz tengan un rango de volumen similar. Esto es útil si vamos a trabajar con múltiples grabaciones que pueden haber sido capturadas en condiciones distintas (grabadas en dispositivos diferentes, etc.). Si no se normalizan las señales de audio, estas diferencias de volumen pueden inducir a errores en el análisis, dificultando las comparaciones de las voces. En el contexto de la detección de voces clonadas, la normalización ayuda a asegurar que el sistema se enfoque en las características en las que se tiene que enfocar de la señal, y no en variaciones de volumen entre muestras en las que este puede variar mucho.

Filtrado de Ruido

El **filtrado de ruido** es otra técnica del preprocesamiento de señales de audio. Muchas veces, las grabaciones de voz contienen ruidos de fondo no deseados, como sonidos o interferencias de equipos cómo el micrófono. Estos ruidos pueden ocultar las características relevantes de la voz. Existen los **filtros de paso bajo** y los **filtros de paso alto**, que son algunas de las herramientas utilizadas para eliminar frecuencias no deseadas de la señal. Por ejemplo, un filtro de paso bajo puede eliminar frecuencias por encima del umbral de la voz humana y al contrario el de paso alto.

Para nuestro sistema, el filtrado de ruido es crucial, ya que las voces generadas artificialmente pueden tener características sutiles que no se aprecian tan fácilmente si tenemos ruido de fondo en las grabaciones. El uso del filtrado debe ser preciso, y saber cuando tenemos que utilizarlo ya que existen situaciones en las que la ausencia total de ruido de fondo puede indicar que la voz ha sido clonada.

Eliminación de Silencios

La **eliminación de silencios** es un método para eliminar los intervalos de inactividad de una grabación de audio. Las pausas frecuentes o duraderas en las grabaciones pueden aumentar la carga computacional, el espacio que ocupan y reducir la eficiencia del sistema, agregando redundancia innecesaria a los datos. El análisis directo de la información que realmente es relevante de la señal se facilita al eliminar las pausas.

Segmentación

La **segmentación** es el proceso de dividir una señal de audio en partes más pequeñas y manejables. Esta técnica puede ser útil para analizar características específicas de la señal en intervalos definidos, detectando alteraciones localizadas que podrían pasar desapercibidas en un análisis

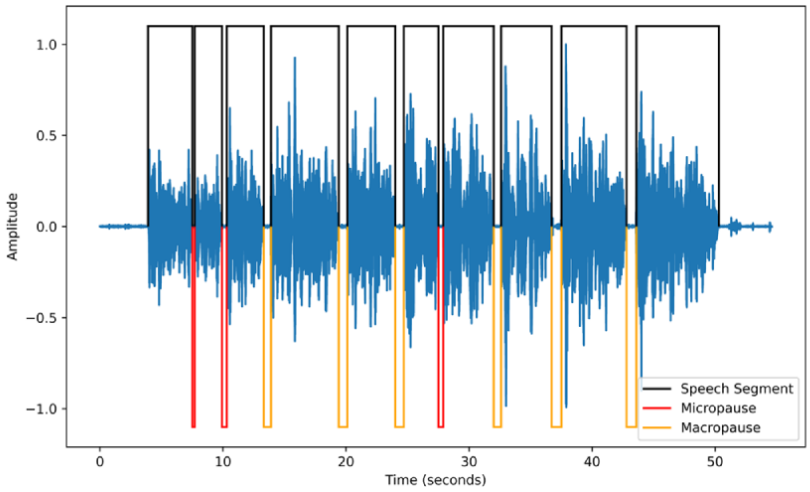


Figura 3.1: Los segmentos negros indican segmentos de habla, los segmentos rojos ilustran micropausas (pausas de menos de 0.5 segundos y de al menos 0.1 segundos), y los segmentos amarillos indican macropausas (pausas de 0.5 segundos o más).[27]

global. En la detección de voces clonadas, la segmentación permite aplicar algoritmos que se especializan en la identificación de patrones en pequeños bloques de audio. Como las manipulaciones en una voz clonada no tiene por que estar distribuidas de manera uniforme en toda la grabación de la voz, sino que pueden ocurrir en momentos puntuales, la segmentación puede mejorar la capacidad del sistema para identificar estos patrones.

Impacto del Preprocesamiento en la Detección de Voces Clonadas

Al eliminar el ruido, los silencios y las variaciones accidentales de volumen, se disminuye la posibilidad de que el sistema interprete como anomalías elementos que no forman parte de la voz. Sin embargo, es importante tener precaución, ya que muchas veces ruidos, silencios, etc... en la señal son indicadores clave de que una voz es real o no. Las voces humanas suelen presentar imperfecciones y pausas que pueden ser eliminadas al limpiar demasiado el audio[27], mientras que las voces clonadas tienden a ser más limpias, lo que puede ayudar en su identificación. Por tanto, aplicar técnicas de preprocesamiento debe equilibrarse para no comprometer al sistema.

3.1.3. Transformación de Señales: Espectrogramas y MFCCs

Los espectrogramas y los Coeficientes Cepstrales de Frecuencia de Mel (MFCCs) permiten realizar un análisis profundo y detallado de las señales de audio.

Espectrogramas: Detección de Patrones Anómalos en la Frecuencia

Un **espectrograma** es una representación visual de cómo las frecuencias de una señal de audio varían con el tiempo[28]. Básicamente, convierte la señal en una imagen, donde el eje horizontal representa el tiempo, el eje vertical las frecuencias, y el color refleja la amplitud/energía en cada

frecuencia a lo largo del tiempo. Esta representación permite detectar cambios o anomalías.

En la detección de voces clonadas, los espectrogramas son útiles porque las voces generadas artificialmente a menudo presentan inconsistencias en las frecuencias altas o bajas. Las redes neuronales profundas pueden entrenarse utilizando espectrogramas, identificando patrones anómalos característicos de voces clonadas con alta precisión. La utilización de espectrogramas como entrada para modelos de aprendizaje automático ha mejorado significativamente la detección de voces sintetizadas mediante inteligencia artificial, dado que los modelos pueden extraer automáticamente características que se podrían pasar por alto.

MFCCs: Coeficientes Cepstrales de Frecuencia de Mel

Los **Coeficientes Cepstrales de Frecuencia de Mel** (MFCCs) son una de las herramientas más utilizadas para la representación y análisis de señales de voz[29]. Representan una transformación de la señal de audio que refleja la percepción humana del sonido, emulando cómo el oído humano responde a las diferentes frecuencias. Esta transformación concentra la mayor parte de la información relevante en las primeras componentes cepstrales, lo que permite una reducción significativa de la dimensionalidad sin perder información crítica sobre el contenido de la señal que tenemos.

El proceso para obtener los MFCCs comienza dividiendo la señal y luego aplicando la **Transformada de Fourier** para pasar al dominio de frecuencia. Después, se aplica una **escala de Mel**, que es una escala logarítmica que refleja cómo el oído humano percibe las frecuencias. Finalmente, se calcula el **cepstrum**, que es esencialmente el espectro de la señal de frecuencia. El resultado final es un conjunto de coeficientes que resumen la forma del espectro a lo largo del tiempo.

En la detección de voces clonadas, los MFCCs permiten analizar con precisión las características acústicas de una voz. En algunos sistemas de detección recientes, los MFCCs son utilizados como características de entrada, proporcionando una representación compacta pero lo suficientemente informativa de las señales de audio para el entrenamiento y la evaluación de los modelos de detección.

Eficacia en la Detección Automatizada de Voces Clonadas

Al combinar estas técnicas con redes neuronales y modelos avanzados de IA, los sistemas de detección pueden analizar grandes volúmenes de datos de audio con una precisión mucho más alta. Las redes convolucionales (**CNNs**), por ejemplo, son especialmente buenas para trabajar con espectrogramas, mientras que los modelos basados en aprendizaje profundo como **LSTMs** pueden utilizar los MFCCs para capturar patrones temporales en la señal.

Capítulo 4

Parámetros de la voz

4.1. Introducción a los Parámetros de la Voz

Los **parámetros acústicos de la voz** son las características del sonido que producimos al hablar y que podemos medir, como la frecuencia, la intensidad, el timbre, la duración, etc. Estos parámetros describen la voz humana, y son esenciales para que podamos encontrar la singularidad en cada voz. En el contexto de la **detección de voces clonadas**, estos parámetros se utilizan para identificar diferencias sutiles entre voces naturales y aquellas generadas o manipuladas mediante IA. Los avances en estas **tecnologías de clonación de voz** han permitido generar clonaciones cada vez más realistas, y por ello aparece la necesidad de analizar estos parámetros en profundidad[30]. Estudiar cómo estos pueden ser replicados o alterados en voces clonadas es necesario para desarrollar y estudiar sistemas que puedan distinguir entre una voz auténtica y una falsificación (que cada vez estarán mejor hechas)[31].

Frecuencia Fundamental (F0) en la Voz

La **frecuencia fundamental (F0)** representa la tasa de vibración de las cuerdas vocales en Hercios (Hz)[32]. Este parámetro determina el tono de una voz, es decir, la percepción de si esta es grave o aguda. La **F0** varía entre individuos debido a factores como el tamaño y la tensión de las cuerdas vocales, el sexo, la edad y el estado emocional, y es un componente clave en la identidad de una voz. Como rango muy general, para los hombres, la **F0** suele oscilar entre 97 y 189 Hz, mientras que en las mujeres se encuentra entre 147 y 262 Hz[33].

La **Frecuencia Fundamental** desempeña un papel central en la producción del habla, ya que la vibración de las cuerdas vocales es la que genera el sonido inicial que luego es modulado por otras estructuras como la lengua o el paladar. La **F0** no solo define el tono de una persona, sino que también es crucial para la entonación y la prosodia, lo que permite diferenciar entre frases interrogativas, afirmativas o transmitir emociones[32]. A estos cambios en la **F0** a lo largo del tiempo se les conoce como *contorno de entonación*, una característica que permite interpretar la

intención de la frase.

Por ejemplo, una oración en un tono más alto al final puede ser una pregunta, mientras que un cambio en el tono puede indicar una afirmación o una emoción como la tristeza. La capacidad de controlar la **Frecuencia Fundamental** es necesaria para una percepción natural del habla.

En la **detección de voces clonadas**, la **Frecuencia Fundamental** es de gran importancia. Puede que las voces generadas carezcan de las variaciones naturales en la **F0** que se producen en contextos reales (fluctuaciones sutiles provocadas por emociones, cansancio, ilusión o incluso el entorno en el que la voz es grabada).

Una clonación de voz que no capture estas variaciones podría sonar mecánica y ser demasiado uniforme su **F0**, lo que podría resultar un indicio. Debemos analizar no solo el tono promedio de la voz, sino también cómo varía a lo largo del tiempo en una conversación. Se puede replicar la **F0** básica de un hablante, pero es más difícil imitar los cambios naturales que ocurren durante una frase o una conversación, como el aumento o disminución de tono asociado con las emociones. Es importante integrar la **frecuencia fundamental** para aumentar la precisión de las detecciones.

Intensidad de la Señal de Voz

La **intensidad** y la **energía** son dos parámetros muy importantes en la señal de voz. La intensidad se refiere a la amplitud de la señal de voz, lo que percibimos como el volumen o el nivel de sonoridad. A nivel físico, se relaciona con la presión que generan las cuerdas vocales.

En la producción del habla, como la mayoría de los parámetros, la intensidad varía de manera natural en función de diversos factores, como las emociones, el estado físico de la persona, el contexto y el entorno[34]. Por ejemplo, cuando una persona está enfadada, su voz tiende a elevarse, mientras que una persona cansada o triste puede hablar con una intensidad más baja. Estas variaciones en la intensidad son parte de la **naturalidad** del habla.

En la clonación de voces, los algoritmos de inteligencia artificial intentan replicar las características naturales del habla de un individuo. Sin embargo, la intensidad es difícil de replicar con precisión, especialmente en contextos naturales donde la variabilidad del entorno y las emociones influencian mucho el habla[30]. Las voces clonadas tienden a mostrar patrones de intensidad más uniformes y pueden fallar en imitar los pequeños cambios de intensidad.

Un sistema de detección debe poder evaluar si los cambios de volumen en la señal de voz siguen un patrón natural o artificial. Las voces generadas artificialmente pueden no reproducir de manera precisa los cambios sutiles en la intensidad que ocurren de manera espontánea mientras se habla, como el aumento gradual de la voz para enfatizar algo o la disminución para introducir una pausa.

El Timbre y su Relación con el Espectro de Frecuencia

El **timbre** es una de las cualidades más distintivas del sonido, y es la característica que nos permite diferenciar dos sonidos que tienen la misma frecuencia fundamental y la misma intensidad. En la voz humana, el timbre es lo que caracteriza las voces de diferentes personas, incluso si están diciendo lo mismo con un tono y volumen similares. El timbre depende de la **composición espectral** del sonido, es decir, de las frecuencias secundarias que acompañan a la frecuencia fundamental (F0) de la voz.

La voz humana no está compuesta únicamente por una frecuencia, sino que incluye la frecuencia fundamental (que corresponde al tono base) y varias frecuencias armónicas o secundarias, que son *múltiplos* de la frecuencia fundamental. La combinación de estas frecuencias forma el **espectro de frecuencia**, y determina el timbre de la voz. El espectro de frecuencia es, por tanto, una representación visual de las distintas frecuencias que componen una señal de audio, y refleja la energía y cómo se distribuye entre los armónicos. Las diferencias en el timbre entre dos personas (o incluso en la misma persona bajo distintas circunstancias) se deben a cómo el tracto vocal, las cuerdas vocales y otros factores moldean este espectro de frecuencia[34]. Por ejemplo, una persona con una cavidad oral más grande podría producir un timbre más profundo, ya que ciertas frecuencias armónicas se notarían más en su voz.

El **procesamiento y análisis del espectro de frecuencia** puede revelar diferencias en cómo se distribuyen las frecuencias armónicas y las características espectrales[35]. Las voces clonadas pueden presentar inconsistencias para reproducir correctamente estas frecuencias, ya que los modelos de síntesis de voz no siempre logran replicar con precisión este tipo de *interacciones* entre las frecuencias. Estas inconsistencias pueden resultar en un timbre que se perciba como menos natural o en una distribución espectral con picos de frecuencia atípicos. Además, mientras que las voces humanas naturales tienden a tener un espectro con transiciones suaves entre las diferentes frecuencias, las voces clonadas pueden mostrar un espectro con discontinuidades más fuertes y no con la suavidad natural.

Los sistemas de clonación de voz también pueden utilizar **filtros** para modelar el timbre, pero esto a menudo introduce artificialidad, haciendo que la voz suene plana o metálica debido a una falta de variabilidad espectral. Un sistema de detección puede comparar los **espectrogramas** de una voz sospechosa con los de una voz original para identificar desviaciones en el timbre[36]. Aunque las tecnologías de clonación de voz han avanzado mucho, todavía tienen como mayor dificultad la de replicar la complejidad de las interacciones entre las distintas armonías que dan forma al timbre de nuestra voz[.]

Formantes de la Voz: Definición y Características

Los **formantes** son bandas de frecuencia que surgen de las propiedades acústicas del tracto vocal humano. Cuando producimos sonidos, las cuerdas vocales generan una vibración que, al

pasar por las cavidades, sufre una modificación que varía según la forma y posición del tracto vocal. Estos formantes corresponden a picos de energía en determinadas frecuencias y son los responsables de dar a los sonidos sus características particulares, como el timbre. Cada persona tiene un patrón de formantes **único** debido a las diferencias físicas en el tracto vocal, lo que hace que la voz de cada individuo sea distinguible y única[34].

La relación que existe entre todas estas frecuencias ayuda a la identificación de sonidos, ya que la ubicación y el comportamiento de los formantes son los que hacen que, por ejemplo, no suene igual la letra *a* que la letra *o*. En el análisis de voz, los formantes reflejan tanto el sonido producido como las características individuales de la voz de la persona.

El papel de los formantes en la producción del habla es crucial, particularmente en el caso de las vocales, que están muy marcadas por sus formantes. El primer formante (**F1**) se asocia con la altura de la lengua respecto al paladar, mientras que el segundo formante (**F2**) refleja el posicionamiento de la lengua dentro de la boca etc...[37] Los formantes no solo nos ayudan a diferenciar vocales, sino que también juegan un papel fundamental en la formación del resto de letras, particularmente en sonidos que dependen de movimientos rápidos y precisos de la lengua.

Los formantes no tienen por qué ser necesariamente estáticos, sino que pueden variar a lo largo del tiempo en un discurso, dependiendo del contenido y del contexto del habla.

En cuanto a la **detección de voces clonadas**, la reproducción fiel de los formantes depende de la exactitud con la que los sistemas de IA puedan modelar y replicar la fisiología humana. En muchos casos, las voces clonadas presentan inconsistencias en las transiciones de los formantes, lo que resulta en una articulación de los sonidos que no coincide con la forma natural en la que la persona lo haría[30]. Estas anomalías pueden manifestarse como pequeños cambios en la frecuencia o en la duración de los formantes, lo cual es detectable mediante un análisis acústico detallado si entrenamos bien la red neuronal.

Debido a su complejidad, los formantes son considerados uno de los indicadores más fiables para identificar voces clonadas, aunque también es más complicado analizarlos correctamente.

La **velocidad del habla** es un parámetro acústico que se refiere a la cantidad de sílabas, palabras o fonemas que un hablante produce en un tiempo determinado, normalmente se mide en sílabas por segundo o palabras por minuto. En situaciones naturales, la velocidad del habla varía de acuerdo con varios factores contextuales como el estado emocional, el tema de conversación, el nivel de fatiga del hablante, entre otras muchas circunstancias, al igual que el resto de parámetros de los que estamos hablando[34].

En situaciones naturales, las personas ajustan su velocidad del habla dependiendo en gran medida del contexto. Por ejemplo, cuando una persona está nerviosa, suele hablar más rápido de lo habitual; mientras que cuando explica algo, su velocidad puede disminuir para ser más clara, o durante una conversación, los hablantes pueden introducir pausas naturales o cambios de ritmo para enfatizar ciertas palabras o ideas, lo que modifica la velocidad.

Estos cambios en la velocidad del habla son difíciles de replicar para los sistemas de clonación de voz. Las **voces clonadas** tienden a mantener una velocidad más uniforme, lo que puede resultar

poco natural, especialmente cuando están se encuentran en conversaciones largas o espontáneas, donde es muy raro que se mantenga una velocidad constante[38]. Además, estos sistemas pueden tener dificultades para recrear las pausas y variaciones en la longitud de las frases o palabras que ocurren de manera natural. Aunque los sistemas de clonación intentan modificar la velocidad del habla para imitar variaciones, una gran ventaja para la detección es que estas modificaciones pueden parecer forzadas o poco sincronizadas con el contenido o el contexto.

Prosodia y Entonación en la Voz Natural y su Relevancia en la Detección de Voces Clonadas

En la comunicación oral, la **prosodia** se compone de elementos como la entonación, el ritmo, la duración de las sílabas y la pausa que transmiten información emocional, estructural o intencional. Estos parámetros tienen un impacto directo en cómo un oyente percibe la autenticidad y la naturalidad de una voz.

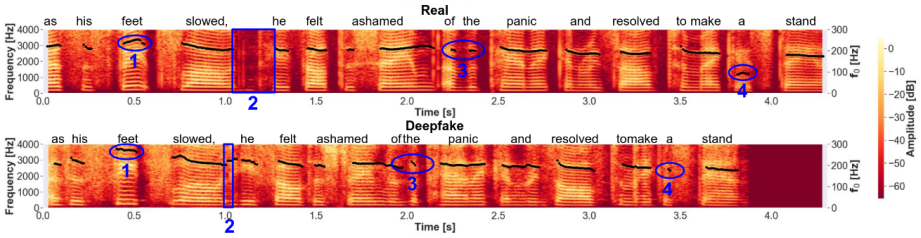


Figura 4.1: Ejemplo de dos espectrogramas. EL primer una voz real, el segundo un modelo entrenado para decir la misma frase[38]

La prosodia y la entonación juegan un papel fundamental en nuestro lenguaje porque reflejan variaciones naturales y contextuales que para los sistemas de clonación de voz basados en IA es lo más difícil de replicar con precisión y naturalidad[38]. Los patrones de prosodia y de entonación de la voz humana son muy variados debido a su constante cambio en función de muchos y distintos factores, lo que también puede dificultar esta detección[39].

La prosodia tiene varias funciones en el habla humana:

- **Estructuración de la Información:** La prosodia organiza la información dentro del discurso. Por ejemplo, una pregunta se suele marcar con una subida en el tono al final de la oración, mientras que una afirmación tiende a tener un descenso en la entonación.

- **Transmisión de Emociones:** El tono y el ritmo son claves para comunicar emociones. Una voz excitada o enfadada mostrará picos de tono más altos y una mayor velocidad de habla, mientras que una voz triste o cansada tendrá una prosodia más plana, con variaciones más suaves en el tono.

- **Intención:** La prosodia ayuda a clarificar la intención del hablante, ya sea para enfatizar una palabra en particular o para señalar una pausa que puede tener un significado dentro del mensaje.

4.2. Métricas de rendimiento utilizadas

Para determinar qué parámetros o distintos conjuntos de parámetros son más óptimos y eficientes para un modelo de detección, se va a entrenar y evaluar el sistema utilizando tanto los atributos por si solos como combinados. Esto permitirá contrastar cómo influye cada característica del sonido en el reconocimiento, buscando tendencias que nos ayuden a que el sistema encuentre el equilibrio necesario entre facilidad para implementarlo, los menos errores posibles y rapidez.

Estas son las métricas que voy a utilizar, basadas en la matriz de confusión[40].

4.2.1. Precisión (Accuracy)

Fórmula:

$$\text{Precisión} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Donde:

- **TP (Verdaderos Correctos):** Voces clonadas correctamente identificadas como clonadas.
- **TN (Negativos Correctos):** Voces originales correctamente identificadas como originales.
- **FP (Falsos Positivos):** Voces originales incorrectamente clasificadas como clonadas.
- **FN (Falsos Negativos):** Voces clonadas incorrectamente clasificadas como originales.

La precisión mide la proporción de predicciones correctas (positivas y negativas) sobre el total de predicciones realizadas. Es útil cómo visión general del rendimiento del modelo, aunque puede ser engañosa en conjuntos de datos que no estén del todo balanceados.

4.2.2. Sensibilidad (Recall)

Fórmula:

$$\text{Sensibilidad} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

La sensibilidad mide la capacidad que tiene el modelo para identificar correctamente todas las voces clonadas. Una alta sensibilidad asegura que el sistema detecta la mayoría de las voces clonadas, y menos falsos negativos. Esta métrica sería adecuada y relevante en sistemas de seguridad, donde es crítico identificar todos los fraudes posibles, aunque se pueda colar cómo clonada alguna voz que no lo es.

4.2.3. Especificidad (Specificity)

Fórmula:

$$\text{Especificidad} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

La especificidad mide la capacidad del modelo para identificar correctamente todas las voces originales. Es fundamental en escenarios donde se necesita minimizar los falsos positivos, por ejemplo, en sistemas de verificación de identidad, donde etiquetar erróneamente una voz real como clonada puede ser un gran problema.

4.2.4. Precisión Positiva

Fórmula:

$$\text{Precisión Positiva} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

La precisión positiva indica qué proporción de las predicciones realizadas como voces clonadas son correctas. Una alta precisión positiva reduce los falsos positivos y es útil cuando se desea maximizar la confianza en las predicciones positivas.

4.2.5. F1-Score[41]

Fórmula:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisión Positiva} \cdot \text{Sensibilidad}}{\text{Precisión Positiva} + \text{Sensibilidad}}$$

El F1-Score es una **media armónica** entre la precisión positiva y la sensibilidad. Es especialmente útil en conjuntos de datos desbalanceados, pues da equilibrio entre la capacidad del modelo para detectar correctamente las voces clonadas y su capacidad para evitar etiquetar incorrectamente las voces originales como clonadas.

4.2.6. AUC-ROC (Área bajo la curva ROC)[42]

Fórmula:

$$\text{AUC} = \int_0^1 \text{Sensibilidad}(1 - \text{Especificidad})\, d(1 - \text{Especificidad})$$

La fórmula se refiere a él área que queda debajo de la curva ROC y representa la probabilidad de que el modelo clasifique correctamente en nuestro caso un audio real y otro clonado elegidos aleatoriamente. Se complementa pues el valor bajo la curva ROC es mejor para una evaluación más general y robusta mientras que F1-Score funciona mejor en escenarios específicos.

4.3. Métricas de eficiencia y rendimiento computacional

4.3.1. Tiempo de Respuesta

Definición: El tiempo de respuesta mide cuánto tarda el modelo en procesar una entrada y generar una predicción.

Fórmula:

$$\text{Tiempo de Respuesta} = T_{\text{final}} - T_{\text{inicio}}$$

Aunque no es la prioridad que el sistema sea rápido si no que sea fiable, las futuras y necesarias adaptaciones al tiempo real no pueden ser posibles si los tiempos de respuesta son elevados.

Capítulo 5

Características y proceso de entrenamiento

5.1. Conjunto de datos utilizados

Para las primeras pruebas he utilizado voces grabadas por mí de compañeros, familiares etc... y que he clonado manualmente utilizando XTTS, un modelo abierto de clonación de voz. En la evaluación final, debido a la cantidad de datos que necesita un modelo así para ser entrenado correctamente y que los resultados sean fiables he utilizado un dataset con voces reales y clonadas ya normalizadas[43].

Este dataset tiene sus limitaciones, pues las muestras de audio solo están en Inglés y son audios dónde las voces representadas no tienen patologías que afecten a la voz, por lo que no podemos generalizar.

5.1.1. Preprocesamiento y normalización de datos

En el dataset que he elegido ya se aplicaron las técnicas de pre-proceso y la normalización de amplitud, muestreo y volumen en todas estas muestras de audio para así evitar posibles sesgos en el entrenamiento.

5.1.2. División en conjuntos de entrenamiento, validación y prueba

La separación de los datos nos permite evaluar mejor cómo va a funcionar el modelo y puede evitar problemas ya comentados cómo el sobreajuste. A continuación se explican los porcentajes elegidos y por qué.

- **Entrenamiento**: En este caso vamos a utilizar un 80 % de los datos para el entrenamiento. El modelo necesita una cantidad suficiente de datos para capturar los patrones más complejos de las voces pero también debe haber datos disponibles para validar y probar. Si el conjunto es muy pequeño tendríamos que aumentar el porcentaje para tener suficientes datos y al revés si es muy grande.

- **Validación**: Representa el 20 % de los datos y este conjunto se va a utilizar para detectar los sobreajustes en datos de entrenamiento (memorizar datos y no aprender patrones) y para ajustar los hiperparametros. Este porcentaje puede disminuir o utilizar otros modos de validación si el conjunto es demasiado pequeño.
- **Prueba**: Se utiliza otra carpeta de audios diferente para la evaluación de rendimiento final del modelo. Estos datos permiten ver la estimación más real del modelo debido a que son datos que no çonoce".

5.2. Características Acústicas Utilizadas

La finalidad de esta selección de características es conocer si hay diferencias estadísticas en los resultados de las pruebas para clasificar las voces entre reales y clonadas buscando un equilibrio entre precisión en la clasificación y eficiencia computacional. Tras investigar las características medibles de la voz he elegido las que considero más útiles. Algunas con resultados muy buenos ya comprobados y otras mas experimentales.

5.2.1. MFCCs (Coeficientes Cepstrales en la Escala Mel)

Los MFCCs son un conjunto de coeficientes que resumen la forma del espectro a lo largo del tiempo. Capturan los matices más importantes del audio[44] y gracias a esto permite identificar diferencias sutiles entre voces reales y voces clonadas, que pueden sonar parecidas a simple vista.

```
if 'mfcc' in parametros:
                mfccs = librosa.feature.mfcc(y=senal_audio, sr=sr, n_mfcc=40)
                features.extend(np.mean(mfccs, axis=1))
```

Este fragmento calcula los MFCCs a partir de la señal de audio (`senal_audio`) con la frecuencia de muestreo decidida (`sr`). Se generan 40 coeficientes por cada ventana de tiempo usando la función `librosa.feature.mfcc`, que convierte la señal en el dominio de frecuencia Mel y luego aplica las transformaciones correspondientes. Como este proceso da lugar a una matriz (coeficientes por ventana), se calcula la media de cada coeficiente a lo largo del tiempo (`np.mean(mfccs, axis=1)`) para resumir la información de toda la señal en un vector, pues la matriz aumentaría mucho la cantidad de datos. Resumirlo así también hace más fácil la comparación con otros audios.

5.2.2. Delta-MFCCs

Los Delta-MFCCs representan los cambios de los MFCCs con el tiempo, es decir, capturan la dinámica de la voz. Si los MFCCs son una foto fija del timbre, los Delta-MFCCs son los que muestran cómo ese timbre evoluciona[45]. Estos funcionan mejor cundo se consigue imitar muy bien el timbre de voz pero no la variación de este.

```
if 'delta_mfcc' in parametros:
                delta = librosa.feature.delta(mfccs)
                features.extend(np.mean(delta, axis=1))
```

Una vez calculados los coeficientes con `librosa.feature.delta`, se calcula la tasa de cambio de esos coeficientes entre frames. Esto da una nueva matriz con la misma dimensión que la de los MFCCs, pero que refleja sus variaciones, y de nuevo, se calcula la media para representar los datos en un único vector.

Tener la matriz completa de los MCFFs y su gradiente sería mucho más preciso pero acabaría con la eficiencia de estos.

5.2.3. Energía (RMS)

El valor cuadrático medio mide el volumen o intensidad promedio de la grabación (la fuerza o intensidad de la voz). Puede ayudar a detectar silencios, pausas, respoiraciones[46] o inconsistencias en la voz.

```
if 'rms' in parametros:
    features.append(np.mean(librosa.feature.rms(y=senal_audio)))
```

`librosa.feature.rms` calcula la energía en cada segmento temporal del audio. La función devuelve un array de valores RMS para cada frame, y `np.mean` los promedia para obtener un único valor representativo de toda la señal.

5.2.4. Relación Armónico-Ruido (HNR)

La HNR mide la claridad de la voz, comparando qué parte de la señal es armónica (voz "limpia") y qué parte es ruido (irregular)[47]. Las voces humanas tienden a tener un patrón armónico natural, mientras que las voces clonadas pueden generar artefactos que se traduzcan en más ruido en la HNR.

```
def extraer_hnr(senal_audio):
    try:
        harmonic = effects.harmonic(y=senal_audio)
        return np.mean(harmonic)
    except:
        return 0.0

if 'hnr' in parametros:
    features.append(extraer_hnr(senal_audio))
```

La función auxiliar `extraer_hnr` utiliza `effects.harmonic` para separar automáticamente los componentes armónicos del ruido. El valor que retorna representa la proporción media entre la energía armónica y la total. Las voces generadas por IA suelen presentar valores más bajos debido a artefactos en la voz.

5.2.5. Formantes

Los formantes son picos en el espectro de frecuencia de la voz que representan las resonancias del tracto vocal. Son esenciales para diferenciar los sonidos vocálicos como 'a', 'e' o 'u'[48].

```
if 'formantes' in parametros and len(senal_audio) < 100000:
    sound = parselmouth.Sound(senal_audio, sr)
    formants = sound.to_formant_burg(max_number_of_formants=3)
    for i in range(1, 4):
        f = formants.get_value_at_time(i, 0.5, "HERTZ")
        features.append(0 if np.isnan(f) or np.isinf(f) else f)
```

El código utiliza la biblioteca Parselmouth para calcular los tres primeros formantes (podemos usar más o menos formantes) mediante el método de Burg(mejor para audios cortos,menor sensibilidad al ruido). Los valores se extraen en el punto medio de la señal y se manejan los casos especiales (NaN/infinito). Las voces sintéticas pueden muestran relaciones anómalas entre estos formantes.

5.2.6. Contraste Espectral[49]

El contraste espectral mide la diferencia entre los picos y valles del espectro de una señal de audio. La utilidad está en el espectro menor que suelen tener las voces clonadas, que tienden a ser más planas por ser artificiales.

```
if 'spectral_contrast' in parametros:
    contrast = librosa.feature.spectral_contrast(y=senal_audio, sr=sr)
    features.extend(np.mean(contrast, axis=1))
```

`librosa.feature.spectral_contrast` divide el espectro en bandas de frecuencia y calcula el contraste para cada una. El código promedia estos valores a lo largo del tiempo para obtener las características más estables. Las voces artificiales suelen mostrar patrones más uniformes en estas bandas.

5.2.7. Ancho de Banda Espectral

El ancho de banda indica cuánto se dispersa la energía de una señal. Un sonido más puro tiende a tener poca dispersión, mientras que uno complejo más.

```
if 'spectral_bandwidth' in parametros:
    features.append(np.mean(librosa.feature.spectral_bandwidth(y=senal_audio, sr=sr)))
```

La función `librosa.feature.spectral_bandwidth` calcula el ancho de banda como la desviación estándar de la distribución espectral. El valor medio resultante da un valor único representativo del ancho de banda espectral medio de toda la señal.

5.2.8. Tasa de Cruces por Cero

La tasa de cruces por cero cuenta cuántas veces la señal pasa de positivo a negativo[46] (o viceversa). Es un indicador de cuánta energía hay en las frecuencias altas.

```
if 'zero_crossing_rate' in parametros:
    features.append(np.mean(librosa.feature.zero_crossing_rate(y=senal_audio)))
```

`librosa.feature.zero_crossing_rate` mide las transiciones por cero en ventanas temporales y se calcula el promedio de estos valores, que pueden ser demasiadas (por ruido artificial) o muy pocas (por filtrado excesivo) comparado con un valor intermedio en voces reales. Junto a la RMS, la tasa de cruces pro cero puede ayudar a localizar respiraciones en la voz.

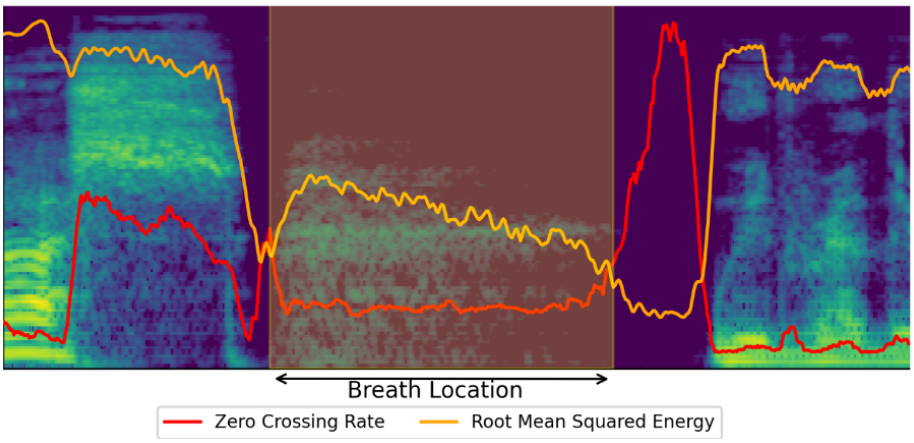


Figura 5.1: Gráfico que muestra la localización de respiración y los valores de RMS y ZCR[46]

5.2.9. Combinaciones Estratégicas

Algunas características, cuando se usan de forma aislada, no ofrecen la suficiente información para detectar con fiabilidad una voz clonada. Por ello, también se plantea la combinación de distintas características, que busca reforzar el sistema de detección aprovechando cómo se complementan entre ellas.

Se han definido varias combinaciones estratégicas:

- **Parámetros individuales:**

- `SOLO_MFCC`: ['mfcc']
- `SOLO_RMS`: ['rms']
- `SOLO_HNR`: ['hnr']
- `SOLO_FORMANTES`: ['formantes']
- `SOLO_SPECTRAL_CONTRAST`: ['spectral_contrast']
- `SOLO_ZERO_CROSSING`: ['zero_crossing_rate']

- **Combinaciones básicas:**

  - `MFCCs+RMS`: Incluye los coeficientes cepstrales MFCC y la energía (RMS). Es una combinación básica pero rápida y combina información del timbre con energía.
  - `MFCCs+HNR`: Sustituye la energía por la HNR, que aporta más información sobre los posibles artefactos en el audio.

- **Combinaciones avanzadas:**

  - `BASICO`: Unimos las tres características de .
  - `CALIDAD_VOCAL`: Incorpora los formantes junto a MFCC y HNR. Esta combinación se centra en los artefactos que pueda tener la voz.
  - `ESPECTRAL_AVANZADO`: Añade contraste espectral y ancho de banda espectral a los MFCC lo que ofrece información más detallada sobre la estructura espectral.

- **Combinación completa:**

  - `ANALISIS_COMPLETO`: Reúne todas las características anteriores. Esta combinación ofrece la máxima precisión en la detección.

Una vez definidas las combinaciones de características acústicas, cada una de ellas será evaluada en una red neuronal cuya arquitectura se describe en la siguiente sección. El objetivo es analizar qué conjunto de características permite una detección más eficaz de voces clonadas. Para comparar el rendimiento de cada combinación, se calculan las métricas habituales en clasificación binaria, como los Verdaderos Positivos (VP), Falsos Positivos (FP), Verdaderos Negativos (VN) y Falsos Negativos (FN), así como indicadores derivados como la , precision, , F1-score... Además, se recogen datos adicionales cómo el número de épocas necesarias para el entrenamiento, el tiempo de procesamiento de los audios, el tiempo de entrenamiento del modelo y el tiempo total del experimento, lo que permite valorar tanto la eficacia como la eficiencia computacional del sistema.

5.3. Diseño Arquitectónico de la Red Neuronal

La arquitectura propuesta es una red neuronal CNN diseñada para adaptarse a las diferentes combinaciones de características acústicas utilizadas en la detección de voces clonadas. Aunque existan otras arquitecturas es preferible la simplicidad de esta para que los resultados se puedan generalizar y que estén lo menos contaminados posible por el uso de una arquitectura muy especial o compleja.

5.3.1. Estructura General del Modelo[50]

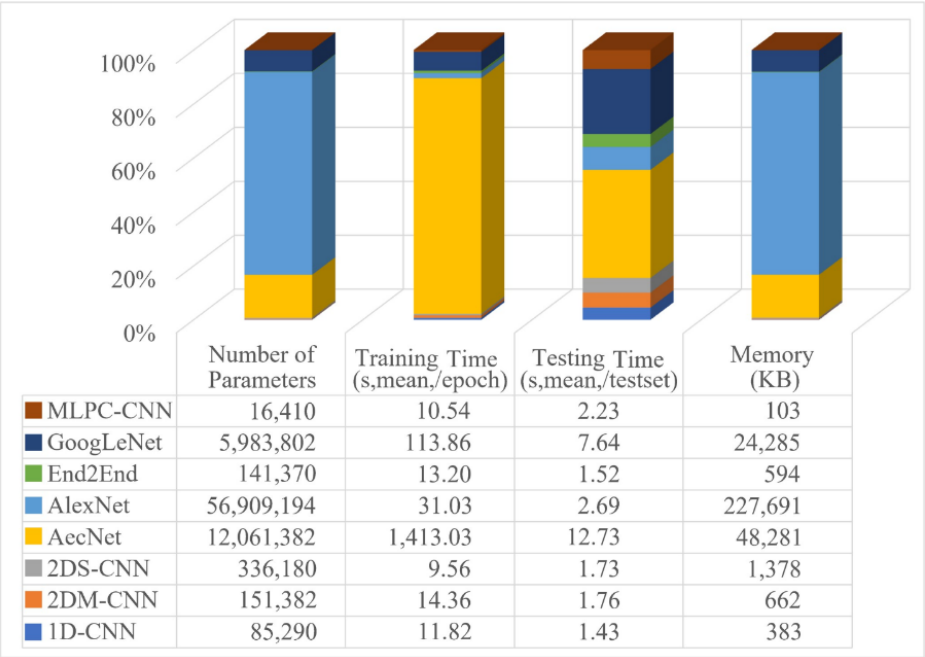
El modelo sigue una estructura secuencial compuesta por las siguientes capas:

- **Capa de entrada**: Se ajusta automáticamente al tamaño del vector de características acústicas.
- **Capa convolucional 1D (opcional)**: Solo se añade si el vector de entrada tiene más de 5 elementos[51] . Esta capa extrae patrones locales entre características acústicas adyacentes y se activa para aprovechar esos patrones(si no no merece la pena activarlo).
- **Capa de agrupamiento (MaxPooling)**: Reduce la dimensionalidad manteniendo las características más relevantes, usando agrupación por pares[52].
- **Dropout**: Desactiva aleatoriamente un porcentaje de neuronas durante el entrenamiento para evitar que el modelo SOLO memorice los datos y así reducir el sobreajuste[53].
- **Flatten**: Aplana la salida para convertirla en un vector que pueda procesarse por una capa densa.
- **Capa densa oculta**: Una capa completamente conectada que permite al modelo aprender combinaciones complejas de las características extraídas.
- **Capa de salida**: Contiene una única neurona con función de activación sigmoide, que produce un valor entre 0 y 1 indicando la probabilidad de que una voz haya sido clonada.

5.3.2. Funcionamiento Adaptativo

El modelo está diseñado para adaptarse a las diferentes longitudes de entrada. Cuando el número de características supera las 5, se aplica una capa convolucional que trata la entrada como una secuencia. En caso contrario, el modelo pasa directamente a las capas densas.

Justificación del diseño El objetivo principal de esta arquitectura no es estudiar su impacto , sino analizar cómo afecta la selección de características al rendimiento y eficiencia del sistema de detección. Por ello, se opta por una estructura básica que se mantiene lo más constante posible. Solo cuando se utilizan combinaciones con mayor número de características (especialmente



	Number of Parameters	Training Time (s,mean,/epoch)	Testing Time (s,mean,/testset)	Memory (KB)
MLPC-CNN	16,410	10.54	2.23	103
GoogLeNet	5,983,802	113.86	7.64	24,285
End2End	141,370	13.20	1.52	594
AlexNet	56,909,194	31.03	2.69	227,691
AecNet	12,061,382	1,413.03	12.73	48,281
2DS-CNN	336,180	9.56	1.73	1,378
2DM-CNN	151,382	14.36	1.76	662
1D-CNN	85,290	11.82	1.43	383

Figura 5.2: Gráfrico que compara una serie de redes neuronales CNN[50]

aquellas que incluyen MFCCs, que generan múltiples coeficientes) se añade una capa convolucional 1D. Esta capa permite aprovechar la estructura secuencial de los MFCCs y extraer patrones locales relevantes entre estos coeficientes consecutivos, sin cambiar radicalmente la arquitectura general. De este modo, podemos acercarnos a asegurar que las mejoras observadas se deban a las características acústicas y no a las diferencias estructurales del modelo.

5.3.3. Configuración

Durante la compilación del modelo se utilizan los siguientes elementos:

- **Optimizador Adam**: Es un optimizador eficiente en el que se define una *tasa de aprendizaje inicial*, Adam lo va ajustando automáticamente durante el entrenamiento en función del historial de gradientes[54].
- **Función de pérdida**: `binary_crossentropy` es la más adecuada para tareas de clasificación binaria como esta, ya que mide la diferencia entre las probabilidades predichas y los valores reales (0 o 1)[55].

5.3.4. Parámetros Ajustables

La arquitectura permite modificar ciertos parámetros[56], así podemos ver los distintos resultados con varias combinaciones de estos y no tener unos resultados para una única combinación. Los valores de estos hiperpárametros deben de tener unos rangos óptimos[57, 58, 59].

- Número de filtros en la capa convolucional (`conv_filters`)

- Número de neuronas en la capa densa (`dense_units`)
- Tasa de desactivación en Dropout (`dropout_rate`)
- Tasa de aprendizaje inicial (`learning_rate`)
- Número máximo de épocas y paciencia para el early stopping[60]

5.3.5. Ventajas de la Arquitectura

- **Flexibilidad**: Permite analizar distintas combinaciones de características acústicas sin necesidad de rediseñar completamente la red.
- **Eficiencia computacional**: Es un modelo relativamente rápido y se puede probar en todo tipo de dispositivos.

5.4. Proceso de Entrenamiento

El proceso de entrenamiento se ha diseñado como una función en el programa en la que se realiza todo el proceso de entrenamiento. A continuación se explica cómo se lleva a cabo esta evaluación.

5.4.1. Carga y Procesamiento de Datos

El sistema comienza con la carga de audios reales y clonados desde las carpetas correspondientes. Para cada conjunto:

- Se extraen las características especificadas en la combinación.
- Se aplica un preprocesamiento consistente para normalizar los datos.
- Se mide el tiempo de procesamiento de forma separada para voces reales y clonadas.

```
X_real, y_real, real_time = cargar_audios_desde_carpeta(rutas['
    train_real'], 1, parametros, porcentaje)
X_falso, y_falso, fake_time = cargar_audios_desde_carpeta(rutas['
    train_fake'], 0, parametros, porcentaje)
X = np.vstack((X_real, X_falso))
y = np.concatenate((y_real, y_falso))
```

Filtro por tamaño: Si el conjunto resultante tiene menos de 10 muestras (no es nuestro caso), se descartaría la combinación por insuficiencia de datos.

5.4.2. Normalización y División de Datos

Los datos se normalizan para tener una convergencia más rápida y estable del modelo:

```
escalador = StandardScaler()
X = escalador.fit_transform(X)
```

A continuación, los datos se dividen para el entrenamiento del modelo y para los test. Se utiliza stratify para mantener la proporción de voces verdaderas/falsas en ambos conjuntos.

```
X_entrenamiento, X_prueba, y_entrenamiento, y_prueba =
   train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=X)
```

5.4.3. Construcción del Modelo

El modelo se construye usando la arquitectura que hemos detallado en la sección anterior.

```
modelo = construir_modelo_cnn((X_entrenamiento.shape[1],),
   config_modelo)
```

5.4.4. Parada Temprana

EarlyStopping detiene el entrenamiento cuando la pérdida en validación no mejora tras varias épocas (patience), y recupera los pesos del modelo cuándo obtuvo mejor rendimiento, evitando sobreajuste y optimizando resultados.

```
parada_temprana = EarlyStopping(
    monitor='val_loss',
    patience=config_modelo['early_stopping_patience'],
    restore_best_weights=True
)
```

5.4.5. Entrenamiento del Modelo

La función `modelo.fit()` ejecuta el proceso de entrenamiento con los siguientes parámetros clave:

- **Épocas**: Número máximo de iteraciones sobre los datos
- **Lotes**: Procesamiento en grupos de 32 muestras
- **Validación**: Reserva automática del 20 % para evaluación
- **Monitorización**: Detiene el entrenamiento si no hay mejora

Durante el entrenamiento, el modelo ajusta sus pesos para minimizar la función de pérdida, almacenando el historial de métricas en cada iteración.

5.4.6. Métricas de Evaluación

El modelo se evalúa mediante un conjunto completo de métricas agrupadas en cuatro categorías:

- **Exactitud**
  - Accuracy (Exactitud global)
  - Precisión (VP / (VP + FP))
  - Recall/Sensibilidad (VP / (VP + FN))
  - F1-Score (Media armónica de precisión y recall)
- **Análisis de curvas**
  - AUC-ROC (Área bajo curva ROC)
  - AUC-PR (Área bajo curva Precisión-Recall)
- **Errores de clasificación**
  - Matriz de confusión:
    - Verdaderos Positivos (VP)
    - Falsos Positivos (FP)
    - Verdaderos Negativos (VN)
    - Falsos Negativos (FN)
- **Eficiencia computacional**
  - Tiempo de procesamiento
  - Tiempo de entrenamiento
  - Tiempo total de ejecución

La evaluación del modelo considera tanto la capacidad predictiva como la eficiencia computacional.

Capítulo 6

Análisis de Resultados

6.1. Marco Experimental

6.1.1. Métricas de Evaluación

Para la evaluación del rendimiento de los modelos de detección se utilizaron las siguientes métricas basadas en la matriz de confusión[40]: Precision, Recall, Specificity, Accuracy, F1-Score y AUC-ROC.

6.1.2. Características Acústicas Analizadas

Parámetros Individuales

Se eligen los siguientes parámetros acústicos que van a ser evaluados de forma individual:

- **MFCCs**: Coeficientes Cepstrales en la Escala Mel (40)
- **RMS**: Energía cuadrática media de la señal
- **HNR**: Relación Armónico-Ruido
- **Formantes**: Primeros tres formantes del tracto vocal
- **Contraste Espectral**: Diferencia entre picos y valles espectrales
- **Tasa de Cruces por Cero**: Frecuencia de cambios de signo en la señal

Combinaciones Estratégicas de Características

Se definen las siguientes combinaciones para análisis comparativo:

1. **MFCCs + RMS**
2. **MFCCs + HNR**

3. **Espectral Avanzado**

4. **Calidad Vocal**

5. **Análisis Total**

6.1.3. Configuraciones Arquitecturales

Se evalúan las características en 20 configuraciones de la red neuronal organizadas en cuatro categorías:

Configuraciones Rápidas (1-5)

Mayor velocidad de procesamiento:

Config	Conv Filters	Dense Units	Dropout	LR	Epochs	Patience
1	16	16	0.1	0.01	10	2
2	32	32	0.2	0.01	15	3
3	64	32	0.2	0.005	20	4
4	32	64	0.3	0.001	15	3
5	16	128	0.4	0.0005	25	5

Configuraciones Balanceadas (6-10)

Equilibrio

Config	Conv Filters	Dense Units	Dropout	LR	Epochs	Patience
6	64	64	0.3	0.001	20	5
7	128	64	0.4	0.0005	25	5
8	64	128	0.4	0.0005	25	5
9	256	64	0.5	0.0001	30	7
10	64	256	0.5	0.0001	30	7

Configuraciones Profundas (11-15)

Más capacidad de aprendizaje:

Config	Conv Filters	Dense Units	Dropout	LR	Epochs	Patience
11	128	128	0.5	0.0001	30	7
12	256	256	0.5	0.0001	40	10
13	512	256	0.6	0.00001	50	15
14	256	512	0.6	0.00001	50	15
15	512	512	0.6	0.00001	50	15

Configuraciones Especializadas (16-20)

Arquitecturas específicas:

Config	Conv Filters	Dense Units	Dropout	LR	Epochs	Patience
16	32	512	0.7	0.0001	40	10
17	512	32	0.7	0.0001	40	10
18	1024	64	0.8	0.00001	60	20
19	64	1024	0.8	0.00001	60	20
20	2048	128	0.9	0.000001	100	30

6.2. Metodología de Análisis Estadístico

6.2.1. Objetivo del Análisis Comparativo

El objetivo principal es determinar si existen diferencias estadísticamente significativas entre las combinaciones de características acústicas y las configuraciones de la arquitectura en el rendimiento de detección de voces clonadas. Esto pretende apoyar la importancia de una buena elección de características con las que entrenar a un modelo independientemente de que tipo de red o los hiperparámetros con los que se entrene y de la dominancia de esta sobre la arquitectura de la red.

6.2.2. Protocolo de Validación Estadística

Verificación de la normalidad de los resultados

1. **Test de Shapiro-Wilk[61][62][63]** ($\alpha = 0{,}05$): Evaluación de normalidad en residuos
    - H_0: Los residuos siguen distribución normal
    - H_1: Los residuos no siguen distribución normal

Tests Estadísticos Principales

Según el resultado de normalidad:

Para datos con residuos normales:

- **ANOVA[64]**: Comparación de medias entre grupos
- Test post-hoc: Tukey HSD[65] para comparaciones múltiples

Para datos con residuos no normales:

- **Test de Kruskal-Wallis[66][67]**: Comparación no paramétrica de distribuciones
- Test post-hoc: Mann-Whitney[68][69] con corrección de Bonferroni[70][71]: La utilizamos pues si mantenemos el valor de alpha sería muy probable encontrar diferencias "significativas" solo por azar debido al elevado número de comparaciones.

6.2.3. Índice de Dominancia

Para cuantificar la importancia relativa :

$$\text{Índice de Dominancia} = \frac{\varepsilon^2_{\text{Combinación}}}{\varepsilon^2_{\text{Arquitectura}}} \tag{6.1}$$

Interpretación:

- $> 10{,}0$: Dominancia muy fuerte de las combinaciones
- $0{,}8 - 1{,}2$: Efectos equivalentes
- $< 0{,}1$: Dominancia muy fuerte de la arquitectura

6.3. Análisis de Resultados por Métricas de Rendimiento

6.3.1. Precision

Comparación entre porcentajes de datos

Cuadro 6.1: Análisis estadístico de Precision por porcentaje de datos

Porcentaje	H-estadístico	ϵ^2_{Comb}	ϵ^2_{Arc}	Índice D	Significancia
1 %	175.98	0.7236	0.0183	39.54	p <0.001
5 %	191.38	0.7911	0.0283	27.99	p <0.001
10 %	204.12	0.8145	0.0245	33.24	p <0.001

Comparaciones post-hoc por porcentaje

Cuadro 6.2: Comparaciones Mann-Whitney significativas para Precision

Porcentaje	Comparaciones totales	Significativas	Tasa éxito
1 %	66	45	68.2 %
5 %	66	42	63.6 %
10 %	66	38	57.6 %

Cuadro 6.3: Estadísticas descriptivas comparativas para Precision (1 %, 5 % y 10 % datos)

Combinación	1 %				5 %				10 %			
	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV
ANALISIS_COMPLETO	0.9419	0.9434	0.0283	0.0301	0.9750	0.9760	0.0136	0.0139	0.9856	0.9861	0.0097	0.0098
ESPECTRAL_AVANZADO	0.9325	0.9354	0.0410	0.0439	0.9720	0.9739	0.0158	0.0162	0.9779	0.9841	0.0170	0.0174
BASICO	0.9180	0.9183	0.0296	0.0322	0.9551	0.9519	0.0233	0.0244	0.9711	0.9791	0.0222	0.0228
CALIDAD_VOCAL	0.9171	0.9302	0.0579	0.0632	0.9502	0.9502	0.0258	0.0271	0.9681	0.9738	0.0235	0.0243
MFCCs+HNR	0.9063	0.9048	0.0371	0.0409	0.9554	0.9622	0.0283	0.0296	0.9640	0.9727	0.0253	0.0262
SOLO_MFCC	0.8976	0.9047	0.0466	0.0519	0.9517	0.9537	0.0251	0.0264	0.9675	0.9724	0.0227	0.0235
MFCCs+RMS	0.8880	0.8798	0.0466	0.0525	0.9514	0.9618	0.0291	0.0306	0.9690	0.9775	0.0254	0.0263
SOLO_SPECTRAL_CONTRAST	0.8039	0.7931	0.0978	0.1216	0.8757	0.8779	0.0479	0.0547	0.8968	0.9132	0.0456	0.0509
SOLO_FORMANTES	0.6679	0.6991	0.1147	0.1718	0.7282	0.7363	0.0697	0.0957	0.7729	0.7760	0.0886	0.1146
SOLO_RMS	0.6629	0.7206	0.1857	0.2801	0.7392	0.7491	0.0330	0.0446	0.7586	0.7652	0.0268	0.0353
SOLO_ZERO_CROSSING	0.5972	0.5946	0.0747	0.1250	0.6335	0.6335	0.0353	0.0556	0.6337	0.6327	0.0275	0.0435
SOLO_HNR	0.5626	0.5567	0.0225	0.0400	0.5616	0.5654	0.0183	0.0326	0.5636	0.5639	0.0065	0.0115

Visualizaciones comparativas

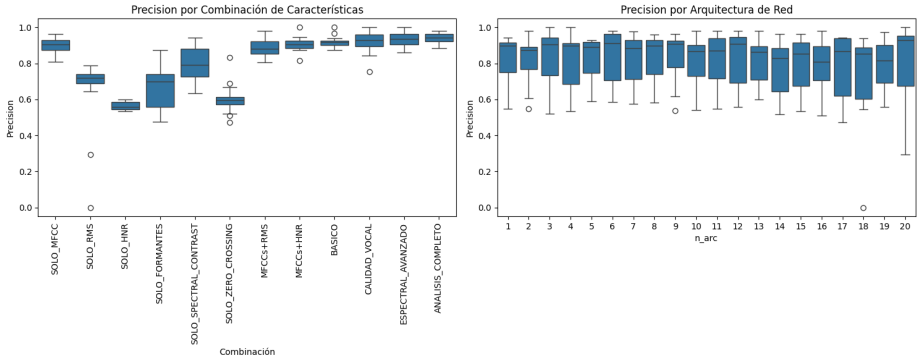


Figura 6.1: Distribución de Precision por combinación de características y arquitectura (1 % datos).

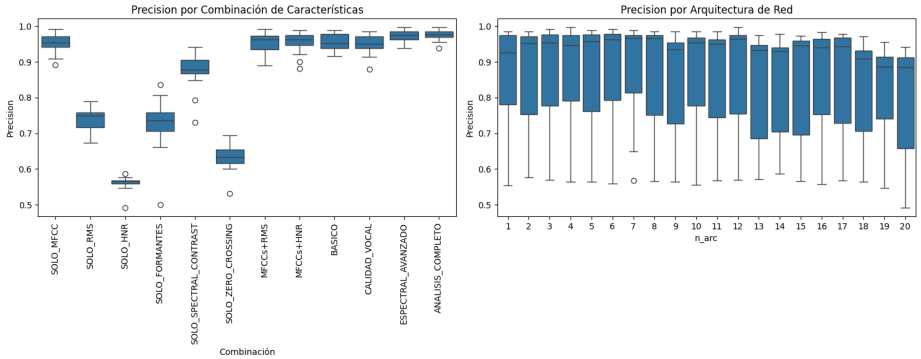


Figura 6.2: Distribución de Precision por combinación de características y arquitectura (5 % datos).

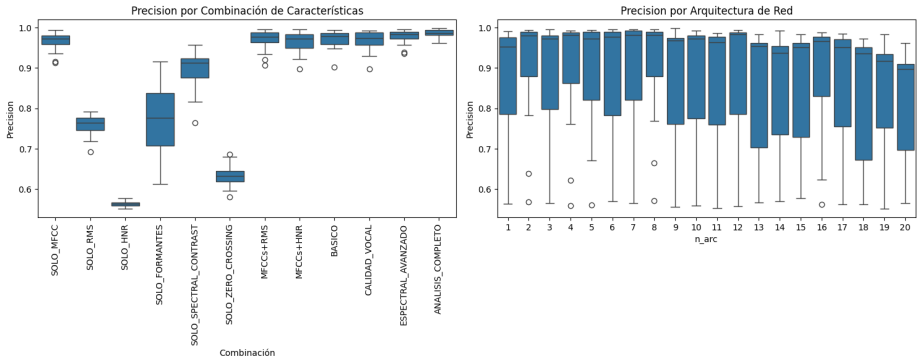


Figura 6.3: Distribución de Precision por combinación de características y arquitectura (10 % datos).

```
TOP 3 - PRECISION
================================================
1. ANALISIS_COMPLETO
   Media global : 0.9675
   CV medio     : 0.0179
   Variabilidad : 0.0228
   Consistencia : Posición promedio #1.0
2. ESPECTRAL_AVANZADO
   Media global : 0.9608
   CV medio     : 0.0259
   Variabilidad : 0.0247
   Consistencia : Posición promedio #2.0
3. BASICO
   Media global : 0.9481
   CV medio     : 0.0265
   Variabilidad : 0.0272
   Consistencia : Posición promedio #3.3
```

Figura 6.4: Top3 conjuntos de características para Precision

6.3.2. Recall

Comparación entre porcentajes de datos

Cuadro 6.4: Análisis estadístico de Recall por porcentaje de datos

Porcentaje	H-estadístico	ϵ^2_{Comb}	ϵ^2_{Arc}	Índice D	Significancia
1 %	168.45	0.6945	0.0156	44.52	p <0.001
5 %	181.20	0.7465	0.0205	36.43	p <0.001
10 %	187.93	0.7892	0.0198	39.86	p <0.001

Comparaciones post-hoc por porcentaje

Cuadro 6.5: Comparaciones Mann-Whitney significativas para Recall

Porcentaje	Comparaciones totales	Significativas	Tasa éxito
1 %	66	43	65.2 %
5 %	66	38	57.6 %
10 %	66	35	53.0 %

Cuadro 6.6: Estadísticas descriptivas comparativas para Recall (1 %, 5 % y 10 % datos)

Combinación	1 %				5 %				10 %			
	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV
SOLO_HNR	0.9519	0.9537	0.0337	0.0354	0.9665	0.9628	0.0116	0.0120	0.9639	0.9629	0.0084	0.0087
ANALISIS_COMPLETO	0.9500	0.9537	0.0346	0.0364	0.9755	0.9759	0.0113	0.0116	0.9832	0.9852	0.0068	0.0070
ESPECTRAL_AVANZADO	0.9259	0.9352	0.0721	0.0779	0.9675	0.9721	0.0230	0.0238	0.9774	0.9777	0.0129	0.0132
BASICO	0.8741	0.8889	0.0928	0.1062	0.9346	0.9424	0.0343	0.0367	0.9517	0.9601	0.0331	0.0348
SOLO_MFCC	0.8713	0.9074	0.1026	0.1177	0.9388	0.9480	0.0383	0.0408	0.9553	0.9610	0.0231	0.0242
MFCCs+HNR	0.8685	0.8889	0.1039	0.1196	0.9335	0.9442	0.0383	0.0411	0.9552	0.9666	0.0319	0.0334
CALIDAD_VOCAL	0.8537	0.8704	0.1176	0.1377	0.9387	0.9461	0.0390	0.0415	0.9540	0.9583	0.0243	0.0255
MFCCs+RMS	0.8380	0.8519	0.1066	0.1272	0.9403	0.9442	0.0392	0.0417	0.9536	0.9647	0.0325	0.0341
SOLO_SPECTRAL_CONTRAST	0.7750	0.8241	0.1229	0.1586	0.8851	0.9052	0.0705	0.0797	0.9128	0.9258	0.0362	0.0397
SOLO_RMS	0.6787	0.7500	0.2305	0.3396	0.7072	0.6914	0.0441	0.0623	0.6901	0.6827	0.0416	0.0602
SOLO_ZERO_CROSSING	0.6009	0.6111	0.1551	0.2581	0.5822	0.5725	0.0332	0.0570	0.5683	0.5686	0.0283	0.0497
SOLO_FORMANTES	0.3796	0.3519	0.1213	0.3195	0.3786	0.3532	0.1421	0.3752	0.3361	0.3349	0.0569	0.1692

Visualizaciones comparativas

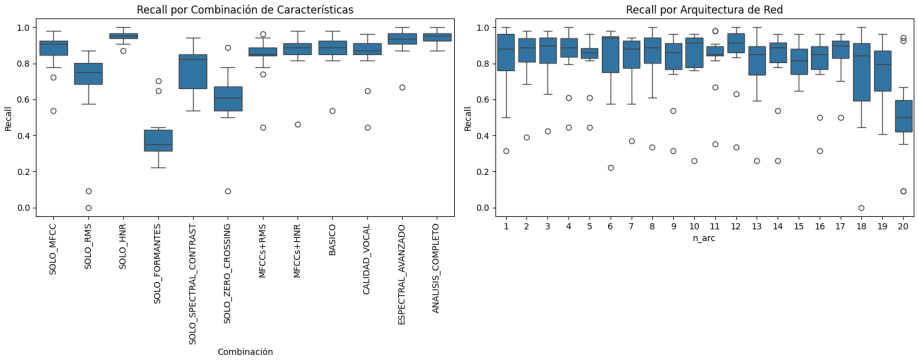


Figura 6.5: Distribución de Recall por combinación de características y arquitectura (1 % datos).

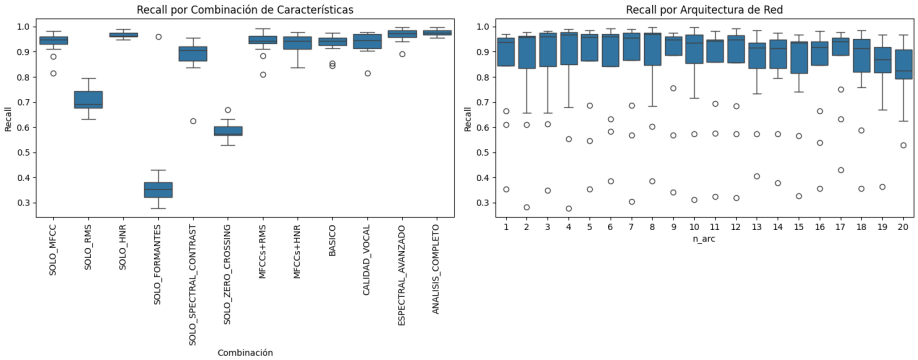


Figura 6.6: Distribución de Recall por combinación de características y arquitectura (5 % datos).

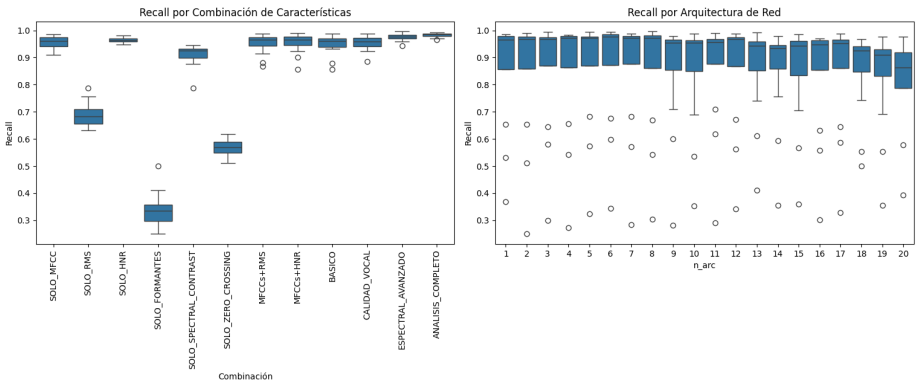


Figura 6.7: Distribución de Recall por combinación de características y arquitectura (10 % datos).

```
TOP 3 - RECALL
==========================================
1. ANALISIS_COMPLETO
   Media global : 0.9696
   CV medio     : 0.0183
   Variabilidad : 0.0174
   Consistencia : Posición promedio #1.3
2. SOLO_HNR
   Media global : 0.9608
   CV medio     : 0.0187
   Variabilidad : 0.0078
   Consistencia : Posición promedio #2.3
3. ESPECTRAL_AVANZADO
   Media global : 0.9569
   CV medio     : 0.0383
   Variabilidad : 0.0273
   Consistencia : Posición promedio #2.3
```

Figura 6.8: Top3 conjuntos de características para Recall

6.3.3. Specificity

Comparación entre porcentajes de datos

Cuadro 6.7: Análisis estadístico de Specificity por porcentaje de datos

Porcentaje	H-estadístico	ϵ^2_{Comb}	ϵ^2_{Arc}	Índice D	Significancia
1 %	152.34	0.6234	0.0289	21.57	p <0.001
5 %	183.79	0.7579	0.0472	16.05	p <0.001
10 %	196.45	0.8156	0.0312	26.14	p <0.001

Comparaciones post-hoc por porcentaje

Cuadro 6.8: Comparaciones Mann-Whitney significativas para Specificity

Porcentaje	Comparaciones totales	Significativas	Tasa éxito
1 %	66	41	62.1 %
5 %	66	44	66.7 %
10 %	66	39	59.1 %

Cuadro 6.9: Estadísticas descriptivas comparativas para Specificity (1 %, 5 % y 10 % datos)

Combinación	1 %				5 %				10 %			
	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV
ANALISIS_COMPLETO	0.9407	0.9444	0.0304	0.0323	0.9750	0.9759	0.0138	0.0142	0.9855	0.9861	0.0098	0.0099
ESPECTRAL_AVANZADO	0.9315	0.9352	0.0446	0.0479	0.9722	0.9741	0.0157	0.0162	0.9777	0.9842	0.0176	0.0180
BASICO	0.9204	0.9259	0.0318	0.0346	0.9563	0.9519	0.0223	0.0233	0.9717	0.9796	0.0212	0.0218
CALIDAD_VOCAL	0.9157	0.9352	0.0707	0.0772	0.9511	0.9500	0.0245	0.0257	0.9685	0.9740	0.0233	0.0240
MFCCs+HNR	0.9065	0.9074	0.0410	0.0452	0.9567	0.9630	0.0272	0.0285	0.9643	0.9730	0.0246	0.0255
SOLO_MFCC	0.8991	0.9074	0.0494	0.0549	0.9526	0.9537	0.0244	0.0257	0.9678	0.9721	0.0228	0.0235
MFCCs+RMS	0.8907	0.8796	0.0527	0.0592	0.9522	0.9630	0.0283	0.0297	0.9695	0.9777	0.0246	0.0254
SOLO_SPECTRAL_CONTRAST	0.8120	0.7870	0.0922	0.1135	0.8754	0.8759	0.0444	0.0508	0.8936	0.9108	0.0501	0.0561
SOLO_FORMANTES	0.7824	0.8426	0.1478	0.1889	0.8343	0.8759	0.1881	0.2255	0.8887	0.9015	0.0739	0.0832
SOLO_RMS	0.7250	0.7222	0.0873	0.1205	0.7478	0.7667	0.0568	0.0760	0.7774	0.7946	0.0458	0.0590
SOLO_ZERO_CROSSING	0.5685	0.5741	0.1549	0.2725	0.6628	0.6611	0.0477	0.0720	0.6692	0.6645	0.0421	0.0629
SOLO_HNR	0.2583	0.2500	0.0599	0.2320	0.2465	0.2574	0.0611	0.2477	0.2520	0.2519	0.0169	0.0672

Visualizaciones comparativas

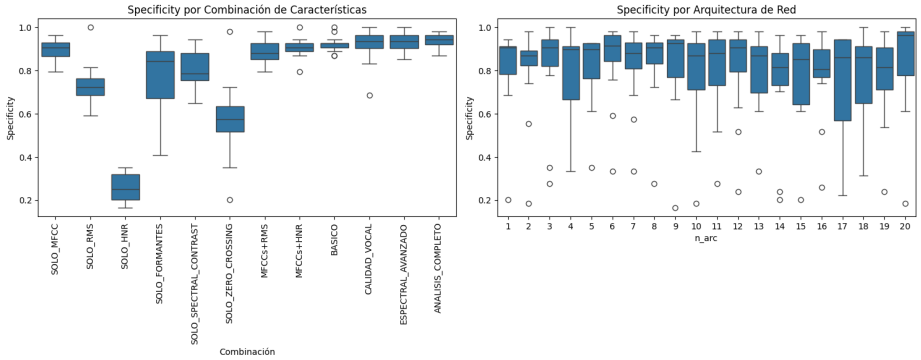


Figura 6.9: Distribución de Specificity por combinación de características y arquitectura (1 % datos).

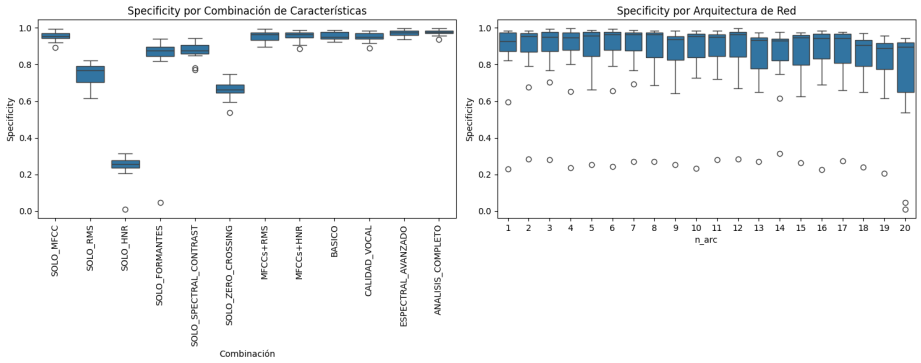


Figura 6.10: Distribución de Specificity por combinación de características y arquitectura (5 % datos).

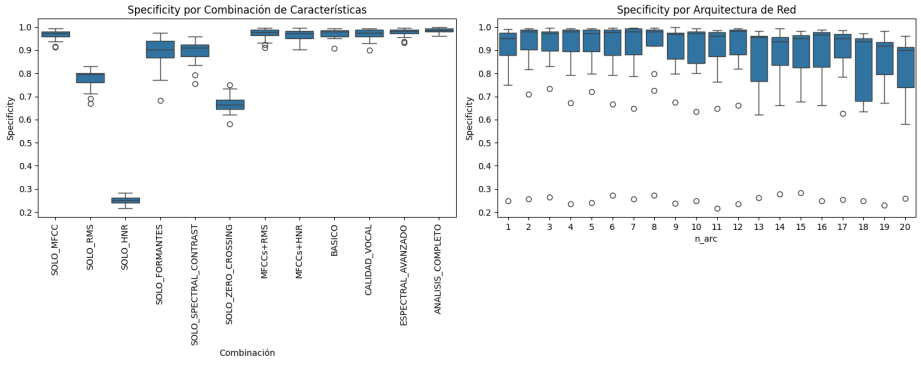


Figura 6.11: Distribución de Specificity por combinación de características y arquitectura (10 % datos).

```
TOP 3 - SPECIFICITY
==========================================
1. ANALISIS_COMPLETO
   Media global : 0.9671
   CV medio     : 0.0188
   Variabilidad : 0.0234
   Consistencia : Posición promedio #1.0
2. ESPECTRAL_AVANZADO
   Media global : 0.9605
   CV medio     : 0.0273
   Variabilidad : 0.0252
   Consistencia : Posición promedio #2.0
3. BASICO
   Media global : 0.9495
   CV medio     : 0.0266
   Variabilidad : 0.0263
   Consistencia : Posición promedio #3.3
```

Figura 6.12: Top3 conjuntos de características para Specificity

6.3.4. Accuracy

Comparación entre porcentajes de datos

Cuadro 6.10: Análisis estadístico de Accuracy por porcentaje de datos

Porcentaje	H-estadístico	ϵ^2_{Comb}	ϵ^2_{Arc}	Índice D	Significancia
1 %	171.23	0.7034	0.0167	42.13	p <0.001
5 %	196.23	0.8124	0.0208	39.03	p <0.001
10 %	203.67	0.8245	0.0223	36.98	p <0.001

Comparaciones post-hoc por porcentaje

Cuadro 6.11: Comparaciones Mann-Whitney significativas para Accuracy

Porcentaje	Comparaciones totales	Significativas	Tasa éxito
1 %	66	42	63.6 %
5 %	66	40	60.6 %
10 %	66	37	56.1 %

Cuadro 6.12: Estadísticas descriptivas comparativas para Accuracy (1 %, 5 % y 10 % datos)

Combinación	1 %				5 %				10 %			
	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV
ANALISIS_COMPLETO	0.9454	0.9398	0.0249	0.0264	0.9752	0.9777	0.0108	0.0110	0.9844	0.9865	0.0070	0.0071
ESPECTRAL_AVANZADO	0.9287	0.9306	0.0428	0.0461	0.9699	0.9722	0.0172	0.0178	0.9775	0.9800	0.0134	0.0137
BASICO	0.8972	0.9074	0.0432	0.0482	0.9455	0.9545	0.0257	0.0272	0.9617	0.9675	0.0257	0.0267
MFCCs+HNR	0.8875	0.8889	0.0450	0.0507	0.9451	0.9499	0.0304	0.0322	0.9597	0.9684	0.0273	0.0285
SOLO_MFCC	0.8852	0.8981	0.0567	0.0641	0.9457	0.9481	0.0284	0.0300	0.9615	0.9684	0.0215	0.0224
CALIDAD_VOCAL	0.8847	0.8935	0.0551	0.0623	0.9449	0.9545	0.0297	0.0314	0.9612	0.9689	0.0231	0.0240
MFCCs+RMS	0.8644	0.8704	0.0507	0.0586	0.9463	0.9536	0.0313	0.0330	0.9616	0.9721	0.0279	0.0290
SOLO_SPECTRAL_CONTRAST	0.7935	0.8148	0.0978	0.1232	0.8802	0.8905	0.0517	0.0587	0.9032	0.9215	0.0394	0.0436
SOLO_RMS	0.7019	0.7269	0.0881	0.1255	0.7276	0.7273	0.0172	0.0236	0.7337	0.7363	0.0115	0.0157
SOLO_HNR	0.6051	0.5972	0.0356	0.0588	0.6058	0.6122	0.0309	0.0511	0.6083	0.6086	0.0104	0.0170
SOLO_ZERO_CROSSING	0.5847	0.5926	0.0491	0.0840	0.6225	0.6206	0.0294	0.0472	0.6187	0.6147	0.0201	0.0325
SOLO_FORMANTES	0.5810	0.5926	0.0443	0.0763	0.6069	0.6104	0.0292	0.0482	0.6121	0.6137	0.0163	0.0266

Visualizaciones comparativas

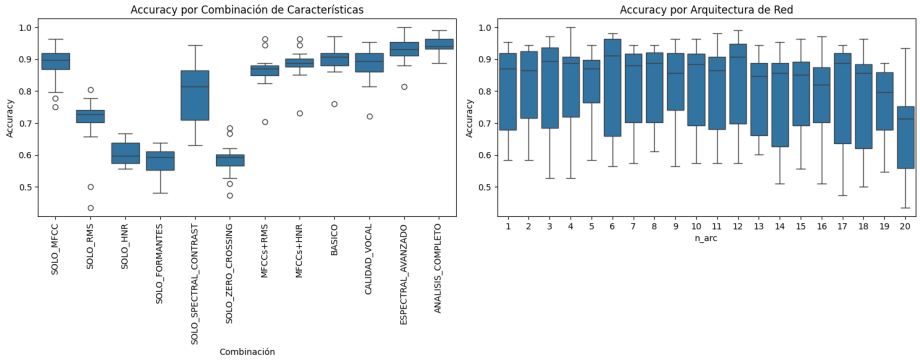


Figura 6.13: Distribución de Accuracy por combinación de características y arquitectura (1 % datos).

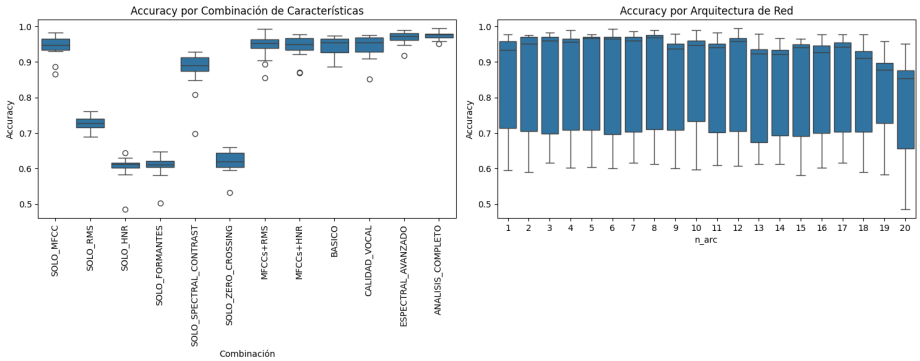


Figura 6.14: Distribución de Accuracy por combinación de características y arquitectura (5 % datos).

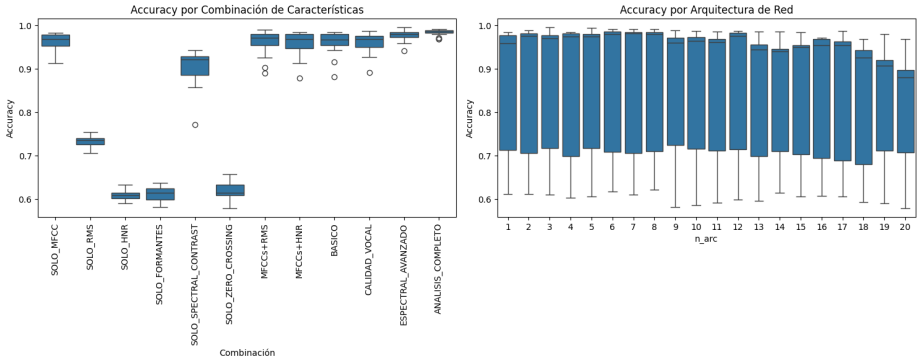


Figura 6.15: Distribución de Accuracy por combinación de características y arquitectura (10 % datos).

```
TOP 3 - ACCURACY
================================================
1. ANALISIS_COMPLETO
   Media global : 0.9683
   CV medio     : 0.0148
   Variabilidad : 0.0204
   Consistencia : Posición promedio #1.0
2. ESPECTRAL_AVANZADO
   Media global : 0.9587
   CV medio     : 0.0258
   Variabilidad : 0.0263
   Consistencia : Posición promedio #2.0
3. BASICO
   Media global : 0.9348
   CV medio     : 0.0340
   Variabilidad : 0.0336
   Consistencia : Posición promedio #3.7
```

Figura 6.16: Top3 conjuntos de características para Accuracy

6.3.5. F1-Score

Comparación entre porcentajes de datos

Cuadro 6.13: Análisis estadístico de F1-Score por porcentaje de datos

Porcentaje	H-estadístico	ϵ^2_{Comb}	ϵ^2_{Arc}	Índice D	Significancia
1 %	173.84	0.7156	0.0171	41.85	p <0.001
5 %	198.52	0.8225	0.0256	32.13	p <0.001
10 %	208.34	0.8387	0.0234	35.84	p <0.001

Comparaciones post-hoc por porcentaje

Cuadro 6.14: Comparaciones Mann-Whitney significativas para F1-Score

Porcentaje	Comparaciones totales	Significativas	Tasa éxito
1 %	66	44	66.7 %
5 %	66	41	62.1 %
10 %	66	36	54.5 %

Cuadro 6.15: Estadísticas descriptivas comparativas para F1-Score (1 %, 5 % y 10 % datos)

Combinación	1 %				5 %				10 %			
	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV
ANALISIS_COMPLETO	0.9455	0.9412	0.0251	0.0265	0.9752	0.9777	0.0107	0.0110	0.9844	0.9866	0.0069	0.0070
ESPECTRAL_AVANZADO	0.9275	0.9296	0.0475	0.0512	0.9697	0.9723	0.0176	0.0182	0.9776	0.9800	0.0132	0.0135
BASICO	0.8922	0.9074	0.0564	0.0633	0.9446	0.9541	0.0268	0.0283	0.9612	0.9673	0.0265	0.0276
MFCCs+HNR	0.8818	0.8878	0.0643	0.0729	0.9442	0.9492	0.0314	0.0333	0.9595	0.9682	0.0278	0.0290
SOLO_MFCC	0.8808	0.8981	0.0679	0.0771	0.9450	0.9482	0.0296	0.0313	0.9613	0.9681	0.0216	0.0224
CALIDAD_VOCAL	0.8770	0.8930	0.0748	0.0853	0.9442	0.9536	0.0308	0.0326	0.9610	0.9685	0.0232	0.0242
MFCCs+RMS	0.8571	0.8670	0.0698	0.0814	0.9457	0.9527	0.0323	0.0342	0.9612	0.9721	0.0285	0.0296
SOLO_SPECTRAL_CONTRAST	0.7874	0.8214	0.1060	0.1347	0.8798	0.8933	0.0563	0.0640	0.9044	0.9222	0.0381	0.0421
SOLO_HNR	0.7069	0.7047	0.0236	0.0334	0.7102	0.7134	0.0158	0.0222	0.7113	0.7116	0.0067	0.0094
SOLO_RMS	0.6656	0.7370	0.2095	0.3147	0.7212	0.7203	0.0175	0.0243	0.7213	0.7206	0.0153	0.0212
SOLO_ZERO_CROSSING	0.5793	0.5992	0.1097	0.1894	0.6062	0.6053	0.0281	0.0463	0.5986	0.5968	0.0197	0.0330
SOLO_FORMANTES	0.4806	0.4673	0.0999	0.2078	0.4808	0.4795	0.0542	0.1127	0.4617	0.4683	0.0382	0.0828

Visualizaciones comparativas

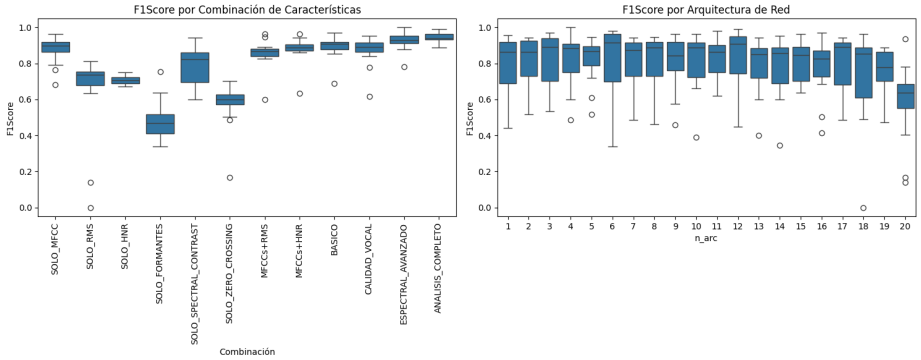


Figura 6.17: Distribución de F1-Score por combinación de características y arquitectura (1 % datos).

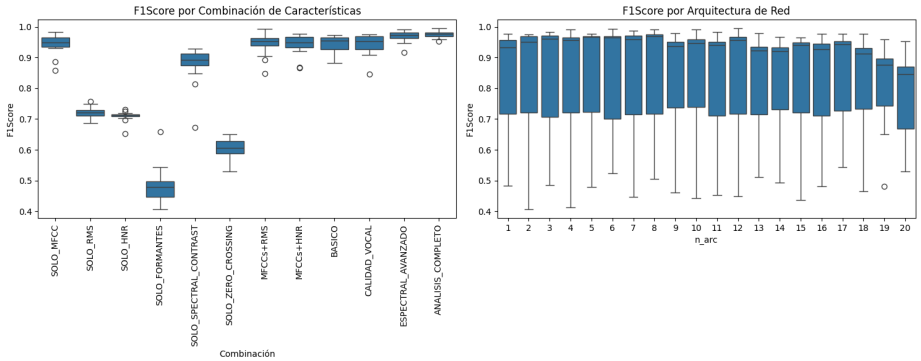


Figura 6.18: Distribución de F1-Score por combinación de características y arquitectura (5 % datos).

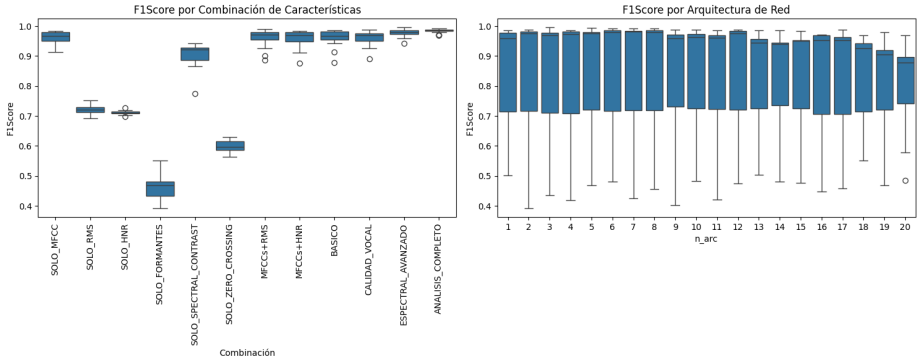


Figura 6.19: Distribución de F1-Score por combinación de características y arquitectura (10 % datos).

```
TOP 3 - F1SCORE
==============================================
1. ANALISIS_COMPLETO
   Media global : 0.9684
   CV medio     : 0.0148
   Variabilidad : 0.0203
   Consistencia : Posición promedio #1.0
2. ESPECTRAL_AVANZADO
   Media global : 0.9583
   CV medio     : 0.0276
   Variabilidad : 0.0269
   Consistencia : Posición promedio #2.0
3. BASICO
   Media global : 0.9327
   CV medio     : 0.0397
   Variabilidad : 0.0360
   Consistencia : Posición promedio #4.0
```

Figura 6.20: Top3 conjuntos de características para F1Score

6.3.6. AUC-ROC

Comparación entre porcentajes de datos

El AUC-ROC presenta los valores más altos de dominancia, siendo la métrica que mejor evidencia la superioridad de las características acústicas.

Cuadro 6.16: Análisis estadístico de AUC-ROC por porcentaje de datos

Porcentaje	**H-estadístico**	ϵ^2_{Comb}	ϵ^2_{Arc}	**Índice D**	**Significancia**
1 %	178.96	0.7389	0.0145	50.96	p <0.001
5 %	204.75	0.8357	0.0189	44.21	p <0.001
10 %	215.23	0.8634	0.0198	43.61	p <0.001

Comparaciones post-hoc por porcentaje

Cuadro 6.17: Comparaciones Mann-Whitney significativas para AUC-ROC

Porcentaje	**Comparaciones totales**	**Significativas**	**Tasa éxito**
1 %	66	47	71.2 %
5 %	66	45	68.2 %
10 %	66	41	62.1 %

Cuadro 6.18: Estadísticas descriptivas comparativas para AUC-ROC (1 %, 5 % y 10 % datos)

Combinación	1 %				5 %				10 %			
	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV	X	Mdn	Desv.	CV
ANALISIS_COMPLETO	0.9858	0.9883	0.0106	0.0107	0.9963	0.9968	0.0026	0.0026	0.9980	0.9986	0.0016	0.0016
ESPECTRAL_AVANZADO	0.9690	0.9743	0.0245	0.0253	0.9931	0.9959	0.0078	0.0079	0.9955	0.9974	0.0053	0.0053
BASICO	0.9631	0.9630	0.0195	0.0202	0.9846	0.9887	0.0124	0.0126	0.9906	0.9945	0.0108	0.0109
CALIDAD_VOCAL	0.9579	0.9628	0.0235	0.0245	0.9854	0.9889	0.0127	0.0129	0.9912	0.9949	0.0087	0.0088
MFCCs+HNR	0.9558	0.9590	0.0209	0.0219	0.9830	0.9871	0.0158	0.0160	0.9899	0.9945	0.0118	0.0120
SOLO_MFCC	0.9512	0.9528	0.0315	0.0332	0.9848	0.9862	0.0133	0.0135	0.9906	0.9952	0.0102	0.0103
MFCCs+RMS	0.9415	0.9424	0.0290	0.0308	0.9841	0.9902	0.0164	0.0167	0.9901	0.9949	0.0114	0.0115
SOLO_SPECTRAL_CONTRAST	0.8895	0.9175	0.0701	0.0788	0.9497	0.9582	0.0373	0.0393	0.9619	0.9672	0.0232	0.0241
SOLO_RMS	0.7637	0.8018	0.1359	0.1779	0.7852	0.7869	0.0184	0.0234	0.7897	0.7911	0.0144	0.0183
SOLO_HNR	0.6292	0.6280	0.0640	0.1016	0.6425	0.6567	0.0720	0.1121	0.6564	0.6593	0.0168	0.0256
SOLO_ZERO_CROSSING	0.6174	0.6158	0.0686	0.1111	0.6499	0.6535	0.0288	0.0443	0.6468	0.6472	0.0269	0.0415
SOLO_FORMANTES	0.5884	0.6037	0.0616	0.1047	0.6364	0.6386	0.0293	0.0460	0.6500	0.6529	0.0254	0.0391

Visualizaciones comparativas

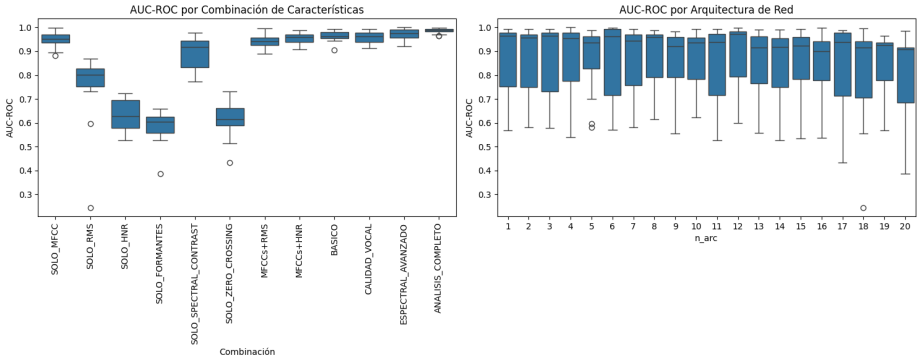


Figura 6.21: Distribución de AUC-ROC por combinación de características y arquitectura (1 % datos).

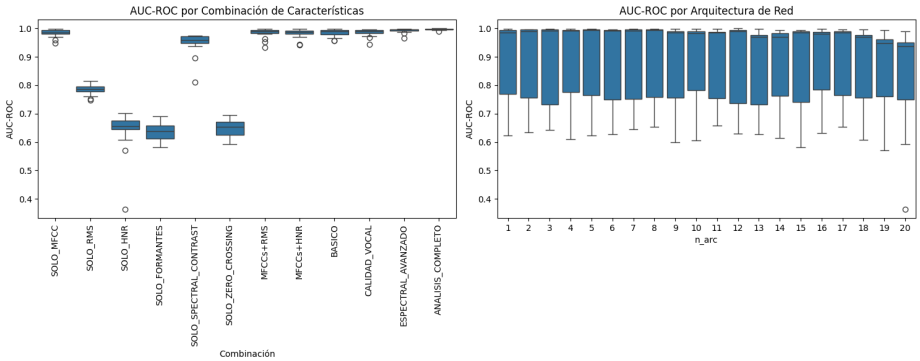


Figura 6.22: Distribución de AUC-ROC por combinación de características y arquitectura (5 % datos).

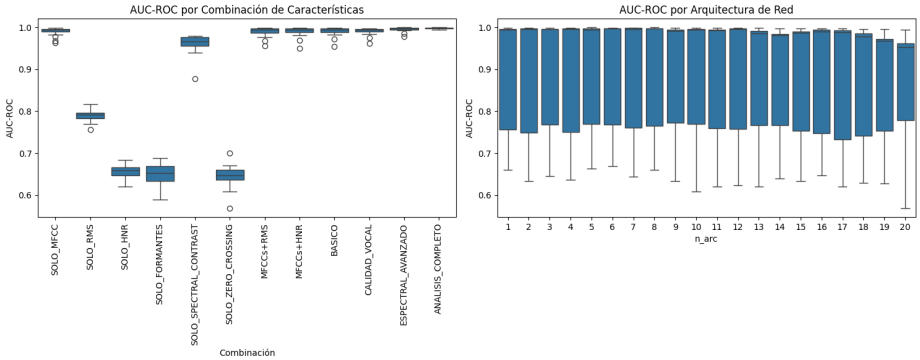


Figura 6.23: Distribución de AUC-ROC por combinación de características y arquitectura (10 % datos).

```
TOP 3 - AUC-ROC
==========================================
1. ANALISIS_COMPLETO
   Media global : 0.9934
   CV medio     : 0.0050
   Variabilidad : 0.0066
   Consistencia : Posición promedio #1.0
2. ESPECTRAL_AVANZADO
   Media global : 0.9859
   CV medio     : 0.0128
   Variabilidad : 0.0147
   Consistencia : Posición promedio #2.0
3. BASICO
   Media global : 0.9794
   CV medio     : 0.0146
   Variabilidad : 0.0145
   Consistencia : Posición promedio #4.0
```

Figura 6.24: Top3 conjuntos de características para AUC-ROC

6.4. Conclusiones

El análisis por métricas revela patrones consistentes en todas ellas que no dependen del tamaño muestral:

1. **Dominancia** : Todas las métricas muestran índices de dominancia muy altos, confirmando la superioridad de las características sobre las arquitecturas.

2. **AUC-ROC es la métrica más óptima**: y presenta los mayores índices de dominancia (43.61-50.96)

3. **Estabilización progresiva**: Los índices de dominancia tienden a estabilizarse con mayor cantidad de datos, reduciendo la variabilidad pero manteniendo la superioridad de las características.

4. **Recall es la métrica más sensible**: Muestra alta variabilidad en función del tamaño muestral, siendo especialmente sensible por lo que es algo menos fiable para generalizar.

Capítulo 7

Mejoras, Limitaciones y Consideraciones Éticas

Este análisis experimental sobre la relevancia de características vocales en la identificación de clonaciones sintéticas ha revelado patrones clave, pero también plantea oportunidades de optimización. A continuación, se exploran las dimensiones críticas para el avance de esta línea de investigación.

7.1. Optimización

7.1.1. Ampliación de Fuentes de Datos

En el estudio actual he utilizado muestras adecuadas para este objetivo pero que pueden ser algo limitadas. Estas han sido grabaciones propias y un dataset ya estandarizado limitado en idioma y patologías vocales. Para poder generalizar los resultados, se necesita tener una base sólida y sería crucial incorporar repositorios masivos. Estos deberían incluir audios en diferentes lenguas, abarcando desde idiomas ampliamente hablados hasta lenguas más minoritarias con características fonéticas únicas. La diversidad lingüística es importante porque diferentes idiomas presentan patrones prosódicos y tonales distintivos que influyen en la efectividad de las características vocales. Además de la diversidad lingüística también se deberían de añadir muestras que incluyan patologías del habla como disfonías, trastornos neurológicos como el Parkinson que afectan la producción vocal, o condiciones como la afasia que alteran los patrones normales del habla para no discriminar a esta parte de la población y que los sistemas sean mas robustos.

7.1.2. Variantes en la arquitectura

Si bien el experimento actual se ha priorizado la simplicidad y usamos redes CNN muy básicas, pues el objetivo era intentar simplificar la detección; pruebas con arquitecturas diferente cómo las transformer(mucho más avanzadas) mostrarían el potencial para capturar dependencias tempora-

les largas que muestra mejores resultados por lo general, superando a los enfoques convolucionales tradicionales. También sistemas que integran redes neuronales especificas para el análisis espectral local y luego redes neuronales recurrentes o mecanismos de atención para capturar patrones temporales de largo plazo darían una comprensión más completa de que características son más distintivas.

7.2. Limitaciones

7.2.1. Avance de la Tecnología

Los sistemas de detección se enfrentan a un avance muy rápido de las técnicas de síntesis, que son cada vez más sofisticadas. Modelos como **NAUTILUS**[72] pueden clonar voces con solo 3 segundos de muestra, incluso generar variaciones prosódicas tan naturales que son indistinguibles para nosotros. Esto tiene un problema, ya que los sistemas más avanzados utilizan cómo indicadores los

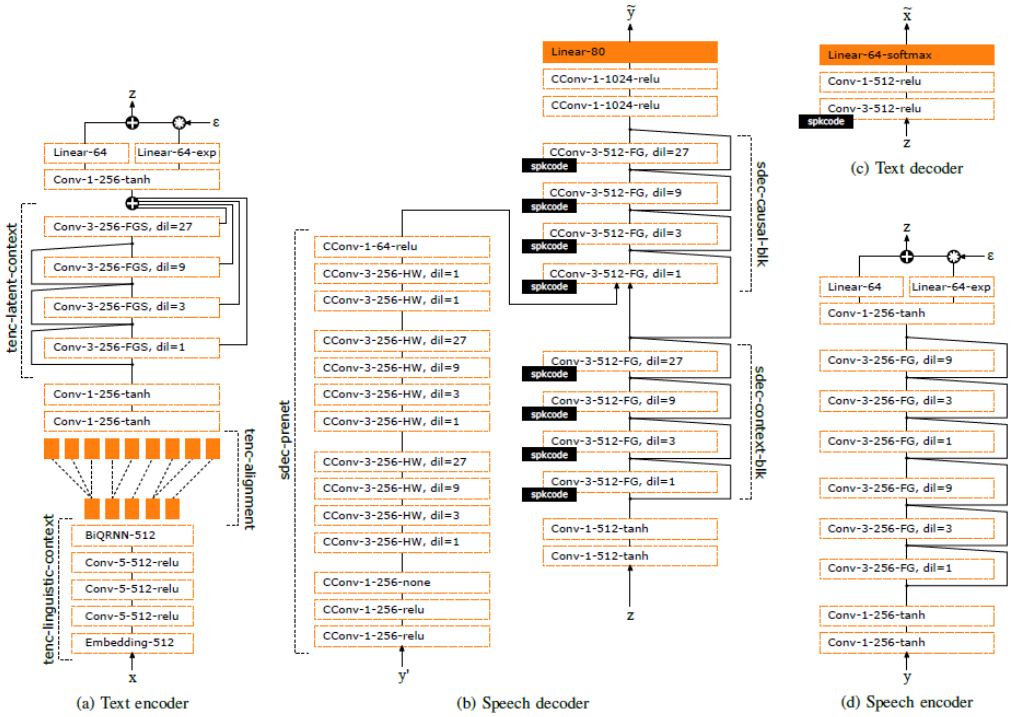


Figura 7.1: Esquema de NAUTILUS[72]

artefactos (que se han considerados indicadores de clonación) y podrían cambiar la características vocales más fiables según vayan mejorando los sistemas de clonación.

La dependencia de la detección a los patrones acústicos hace que resulte vulnerable ante ataques diseñados específicamente para confundir a los algoritmos de detección , por ejemplo insertando perturbaciones imperceptibles. Existen experimentos que han dado lugar a modelos complejos que ayuden a detectar estas perturbaciones o deepfakes[73].

7.2.2. Errores

- Ningún modelo, por avanzado y bien entrenado que esté o utilizando cualquier característica, va a eliminar completamente los errores como Falsos Negativos o Falsos Positivos. Aunque se pueda reducir mucho la tasa de errores, la eliminación completa es un objetivo inalcanzable debido a la naturaleza de los algoritmos de aprendizaje automático. Además, en contextos de alta seguridad, hasta unas tasas muy bajas pueden ser preocupantes debido a la gran cantidad de transacciones que ocurren.

- Existen otros problemas como la generalización específica que pueden llevar a una mayor tasa de error. Por ejemplo, cuando se entrena un modelo en un solo idioma, el modelo tiende a capturar patrones acústicos muy específicos de esa población de entrenamiento, lo que afecta negativamente a su rendimiento cuando se traslada a otros idiomas. Existen técnicas recientes para intentar mejorar esto, incluyendo representación sustancial de idiomas tonales o arquitecturas que puedan adaptarse dinámicamente a diferentes características lingüísticas [74].

7.3. Ética del experimento

Aunque la universalización de la clonación de voces ofrece oportunidades muy buenas en ámbitos como la accesibilidad o la educación, también presenta problemas que no pueden ignorarse. La facilidad creciente para acceder a herramientas de síntesis vocal plantea interrogantes sobre la autenticidad de las comunicaciones y la integridad de la identidad personal en entornos digitales y experimentos como este que analizan las características vocales para detectar diferencias estadísticas entre voces auténticas y clonadas pueden servir como una base científica sólida para la creación de sistemas de detección robustos que protejan contra fraudes y suplantaciones. Cuanto mejores sean estos sistemas de detección, mayor será la posibilidad de que disuadan de utilizar la tecnología de clonación para usos maliciosos. Cualquier implementación necesita estar regulada y estar equilibrada con la protección y la privacidad[75].

Bibliografía

[1] James L. Flanagan. «Speech Analysis, Synthesis and Perception». En: Springer, 1972. Cap. 11, págs. 285-338. URL: https://jontallen.ece.illinois.edu/uploads/537.F18/Book/main-all.pdf.

[2] Jonathan Shen et al. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. 2018. arXiv: 1712.05884 [cs.CL]. URL: https://arxiv.org/abs/1712.05884.

[3] Aaron van den Oord et al. *WaveNet: A Generative Model for Raw Audio*. 2016. arXiv: 1609.03499 [cs.SD]. URL: https://arxiv.org/abs/1609.03499.

[4] William Bogren y Mohamed Abdul Hussein. «Social Media's Take on Deepfakes: Ethical Concerns in the Public Discourse». En: (2024). URL: https://uu.diva-portal.org/smash/get/diva2:1837417/FULLTEXT01.pdf.

[5] R. Traboulsi. «Impact of Deepfake Technology on Social Media: Detection, Societal Implications, and Policy Responses». En: *Journal of Emerging Practices in STEM Education* (2023). URL: http://www.epstem.net/tr/download/article-file/3456697.

[6] Robin San Roman et al. «Proactive Detection of Voice Cloning with Localized Watermarking». En: *arXiv preprint arXiv:2401.17264* (2024). URL: https://arxiv.org/abs/2401.17264.

[7] *Ciberamenazas contra entornos empresariales: Una guía de aproximación para el empresario*. Inf. téc. Sección: Fraude del CEO (spear phishing). Instituto Nacional de Ciberseguridad (INCIBE), 2020, págs. 30-33. URL: https://www.incibe.es/sites/default/files/contenidos/guias/doc/ciberamenazas_contra_entornos_empresariales.pdf.

[8] World Economic Forum. *Cybercrime: Lessons learned from a $25m deepfake attack*. Fraudsters used an AI deepfake to steal $25 million from UK engineering firm Arup in 2024. 2025. URL: https://www.weforum.org/stories/2025/02/deepfake-ai-cybercrime-arup/.

[9] Trend Micro. *Unusual CEO Fraud via Deepfake Audio Steals US$243,000 From UK Company*. Caso real de fraude empresarial con deepfake de voz. 2019. URL: https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/unusual-ceo-fraud-via-deepfake-audio-steals-us-243-000-from-u-k-company.

[10] Joseph Faber. *Euphonia de Joseph Faber*. Máquina de síntesis con cabeza antropomórfica. 1846. URL: https://commons.wikimedia.org/wiki/File:Joseph_Faber%E2%80%99s_%E2%80%9CEuphonia%E2%80%9D_talking_machine.jpg.

[11] Thomas Styger y Eric Keller. «Formant synthesis». En: *Fundamentals of speech synthesis and speech recognition: Basic concepts, state of the art, and future challenges* (1994), págs. 109-128. URL: https://www.erickeller.ch/pdf.files/Styger-Keller-94-FundVol.pdf.

[12] Homer Dudley. *The Carrier Nature of Speech*. Inf. téc. 4. Esquemas técnicos originales del Voder. Bell System Technical Journal, 1940, pág. 509. URL: https://archive.org/details/bellsystemtechni19amerrich/page/509/mode/1up?view=theater.

[13] K. Palacio, J. Auquilla y E. Calle. «Diseño e Implementación de un Sistema de Síntesis de Voz». En: *Revista Tecnológica ESPOL* (2013). Describe la síntesis por concatenación de unidades, el uso de fragmentos pregrabados y las ventajas respecto a métodos anteriores. URL: https://rte.espol.edu.ec/index.php/tecnologica/article/download/145/89.

[14] IBM Research Team. «Text-to-Speech: Components and Evolution». En: (2024). Detalla el proceso de análisis lingüístico y síntesis vocal, incluyendo modelos de redes neuronales profundas y evolución histórica desde sistemas mecánicos hasta IA. Incluye diagramas de flujo del proceso TTS. URL: https://www.ibm.com/think/topics/text-to-speech.

[15] Xu Tan et al. *A Survey on Neural Speech Synthesis*. 2021. arXiv: 2106.15561 [eess.AS]. URL: https://arxiv.org/abs/2106.15561.

[16] Ian Goodfellow, Yoshua Bengio y Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[17] Charu C. Aggarwal. *Neural Networks and Deep Learning: A Textbook*. Springer, 2018, Capitulos 7-9. URL: https://www.springer.com/gp/book/9783319944623.

[18] Simon J.D. Prince. *Understanding Deep Learning*. The MIT Press, 2023, Capitulos 5, 6, 9. URL: http://udlbook.com.

[19] Amazon Web Services. «¿Qué es una red neuronal?» En: *AWS Documentation* (2022). Explicación didáctica y visual de la estructura de una red neuronal artificial: capas de entrada, ocultas y salida, funcionamiento de las neuronas artificiales y analogía con el cerebro humano. URL: https://aws.amazon.com/es/what-is/neural-network/.

[20] Judith E. Dayhoff. *Neural network architectures: an introduction*. USA: Van Nostrand Reinhold Co., 1990. ISBN: 0442207441.

[21] Wenpeng Yin et al. *Comparative Study of CNN and RNN for Natural Language Processing*. 2017. arXiv: 1702.01923 [cs.CL]. URL: https://arxiv.org/abs/1702.01923.

[22] Sepp Hochreiter y Jürgen Schmidhuber. «Long short-term memory». En: *Neural Computation* 9.8 (1997). Describe la arquitectura LSTM, solución al problema del desvanecimiento del gradiente en RNN y base de los modelos secuenciales modernos., págs. 1735-1780. URL: https://www.bioinf.jku.at/publications/older/2604.pdf.

[23] Alan V Oppenheim, Ronald W Schafer, John R Buck et al. «Tratamiento de señales en tiempo discreto». En: (2011), Capitulos 2-4-7. URL: https://d1wqtxts1xzle7.cloudfront.net/64145456/Tratamiento_de_senales_en_tiempo_discreto_3ed_-_Oppenheim-libre.pdf?1597118568=&response-content-disposition=inline%3B+filename%3DTratamiento_de_senales_en_tiempo_discret.pdf&Expires=1749395628&Signature=F8~qeKlStka5eSSWPKcMQMPArQGKEmz-bMKDH2OHockw10~mJNKutSNKmsG9o-kfoL0ty~0-EDUXZpM9HW~PJgWYOjA8DNkUUmnv6pQKadQ5yXc-029v0cBv4WtklCZdF4iE3SqFx6Sqxo2DAiD~yDm9QmPTg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.

[24] Lawrence R. Rabiner y Biing Hwang Juang. *Fundamentals of speech recognition*. eng. Prentice Hall signal processing series. Englewood Cliffs, New Jersey: PTR Prentice - Hall, 1993. ISBN: 0130151572.

[25] John C Glover, Victor Lazzarini y Joseph Timoney. «Python for audio signal processing». En: (2011). URL: https://mural.maynoothuniversity.ie/id/eprint/4115/.

[26] Wikipedia contributors. «Audio normalization». En: *Wikipedia, The Free Encyclopedia* (2024). URL: https://en.wikipedia.org/wiki/Audio_normalization.

[27] Nikhil Valsan Kulangareth et al. «Investigation of Deepfake Voice Detection Using Speech Pause Patterns: Algorithm Development and Validation». En: *JMIR Biomedical Engineering* 9 (2024), e56245. DOI: 10.2196/56245. URL: https://biomedeng.jmir.org/2024/1/e56245.

[28] Zulfidin Khodzhaev. *A Practical Guide to Spectrogram Analysis for Audio Signal Processing*. 2024. arXiv: 2403.09321 [cs.SD]. URL: https://arxiv.org/abs/2403.09321.

[29] Wikipedia contributors. *Mel-frequency cepstrum*. URL: https://en.wikipedia.org/wiki/Mel-frequency_cepstrum.

[30] Albérick Euraste Djiré et al. «Evaluating Acoustic Parameters for DeepFake Audio Identification». En: *2023 IEEE Afro-Mediterranean Conference on Artificial Intelligence (AMCAI)*. 2023, págs. 1-6. DOI: 10.1109/AMCAI59331.2023.10431521.

[31] Daniel Calderón-González et al. «Deep Speech Synthesis and Its Implications for News Verification: Lessons Learned in the RTVE-UGR Chair». En: *Applied Sciences* 14.21 (2024), pág. 9916.

[32] C. García y Daniel Tapias Merino. *La frecuencia fundamental de la voz y sus efectos en el reconocimiento de habla continua*. 2000-09. URL: http://hdl.handle.net/10045/1891.

[33] G. Paolini, A. Hernández y V. Pereyra. «Frecuencia fundamental de habla de voz normal según sexo en la Provincia de Córdoba, Argentina». En: *Revista de la Facultad de Ciencias Médicas de la Universidad Nacional de Córdoba* (2 de oct. de 2018). Consultado el 14 de junio de 2025, págs. 99-100. URL: https://revistas.unc.edu.ar/index.php/med/article/view/21300.

[34] Adriana Moreno Méndez et al. «Parámetros acústicos de la voz en el adulto mayor». En: *Umbral científico* 17 (2010), págs. 9-17. URL: https://www.redalyc.org/pdf/304/30421294002.pdf.

[35] Jhon Jimenez PeÃ, Fernando AarÃ Torres Castillo y Oscar Esaul Cueva Sanchez. «ComparaciÃforense de voces: un estudio preliminar sobre las diferencias entre una voz natural y una voz artificial para la investigaciÃjudicial». es. En: *Revista Oficial del Poder Judicial* 16 (ene. de 2024), págs. 53-81. ISSN: 2663-9130. URL: http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2663-91302024000100053&nrm=iso.

[36] Anagha Sonawane y M. U Inamdar. «Synthetic Speech Spoofing Detection using MFCC and SVM». En: *International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT)* 3.3 (2017). ISSN: 2454-132X, Impact factor: 4.295 (2017), págs. 738-742. URL: https://www.ijariit.com/manuscripts/v3i3/V3I3-1433.pdf.

[37] M. Pilar Murtró Ayats. *Bases Acústicas de la Voz*. https://fonologos.com/nweb/wp-content/uploads/2019/04/Bases-Acusticas-Voz.pdf. Ponencia al XVI Congreso Nacional de la Sociedad Médica Española de Foniatría (SOMEF). 2019. URL: https://fonologos.com/nweb/wp-content/uploads/2019/04/Bases-Acusticas-Voz.pdf.

[38] Kevin Warren et al. *Pitch Imperfect: Detecting Audio Deepfakes Through Acoustic Prosodic Analysis*. 2025. arXiv: 2502.14726 [cs.SD]. URL: https://arxiv.org/abs/2502.14726.

[39] Sheila Queralt y Javier Huertas. *Cuidado si utilizas detectores para saber si un audio está generado por IA*. 2025. URL: https://maldita.es/malditatecnologia/20250603/herramientas-verificacion-audio-inteligencia-artificial/.

[40] IBM. «¿Qué es una matriz de confusión?» En: (2024). Explicación detallada de la matriz de confusión y sus componentes. URL: https://www.ibm.com/es-es/think/topics/confusion-matrix.

[41] Applied AI Course. *F1 Score in Machine Learning*. 2024. URL: https://www.appliedaicourse.com/blog/f1-score-in-machine-learning/.

[42] Arya y otros Tafvizi. «Attributing AUC-ROC to Analyze Binary Classifier Performance». En: *arXiv preprint* (2022). eprint: 2205.11781. URL: https://arxiv.org/abs/2205.11781.

[43] Mohammed Abdeldayem. *The Fake-or-Real (FoR) Dataset (deepfake audio)*. https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset. Kaggle. Version: for-norm, for-original, for-2sec, for-rerec. GNU Lesser General Public License 3.0. Accessed: 2025-06-30. 2025.

[44] Antonio Alexandre Lima, Marcello Montillo Provenza y Maria Augusta S. N. Nunes. «Comics as a Pedagogical Tool for Teaching». En: *2022 XVII Latin American Conference on Learning Technologies (LACLO)*. 2022, págs. 1-7. DOI: 10.1109/LACLO56648.2022.10013316.

[45] Joel Frank y Lea Schönherr. «WaveFake: A Data Set to Facilitate Audio DeepFake Detection». En: *Proc. Neural Information Processing Systems Track on Datasets and Benchmarks*. 2021. URL: https://openreview.net/pdf?id=IO7jcf63iDI.

[46] Rita Yang et al. «Every Breath You Don't Take: Deepfake Speech Detection Using Breath». En: *arXiv preprint arXiv:2404.15143* (2024). URL: https://arxiv.org/abs/2404.15143.

[47] Anuwat Chaiwongyen et al. «Deepfake-speech Detection with Pathological Features and Multilayer Perceptron Neural Network». En: *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2023, págs. 2182-2188. DOI: 10.1109/APSIPAASC58517.2023.10317331. URL: https://www.nict.go.jp/en/asean_ivo/gden8c00000004el-att/pdnf1l00000004ja.pdf.

[48] Luca Cuccovillo, Milica Gerhardt y Patrick Aichroth. «Audio Transformer for Synthetic Speech Detection via Multi-Formant Analysis». En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. Jun. de 2024, págs. 4409-4417. URL: https://openaccess.thecvf.com/content/CVPR2024W/WMF/papers/Cuccovillo_Audio_Transformer_for_Synthetic_Speech_Detection_via_Multi-Formant_Analysis_CVPRW_2024_paper.pdf.

[49] Pranita Palsapure et al. «Discriminative deep learning based hybrid spectro-temporal features for synthetic voice spoofing detection». En: *IAES International Journal of Artificial Intelligence (IJ-AI)* 14.1 (2025), págs. 130-141. ISSN: 2252-8938. DOI: 10.11591/ijai.v14.i1.pp130-141. URL: https://ijai.iaescore.com/index.php/IJAI/article/view/24424.

[50] Q. B. Diep, H. U. Y. Phan y T. C. Truong. «Crossmixed convolutional neural network for digital speech recognition». En: *PloS one* 19.4 (2024), e0302394. DOI: 10.1371/journal.pone.0302394. URL: https://doi.org/10.1371/journal.pone.0302394.

[51] Ali Ehteshami Bejnordi y Ralf Krestel. «Dynamic Channel and Layer Gating in Convolutional Neural Networks». En: *Proceedings of the 28th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. 2020. URL: https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/PDFs/2020_bejnordi_dynamic.pdf.

[52] KDnuggets Team. «Diving into the Pool: Unraveling the Magic of CNN Pooling Layers». En: *KDnuggets* (2025). URL: https://www.kdnuggets.com/diving-into-the-pool-unraveling-the-magic-of-cnn-pooling-layers.

[53] Kashish Gandhi et al. «A Multimodal Framework for DeepFake Detection». En: *arXiv preprint arXiv:2410.03487* (2024). URL: https://arxiv.org/html/2410.03487v1.

[54] DigitalOcean Team. *Audio Classification with Deep Learning*. 2025. URL: https://www.digitalocean.com/community/tutorials/audio-classification-with-deep-learning.

[55] Data Science Stack Exchange Community. «What makes binary cross entropy a better choice for binary classification than other loss functions?» En: (2023). URL: https://datascience.stackexchange.com/questions/53400.

[56] IBM. *¿Qué es el ajuste de hiperparámetros?* 2024. URL: https://www.ibm.com/es-es/think/topics/hyperparameter-tuning.

[57] Mohaimenul Azam Khan Raiaan et al. «A systematic review of hyperparameter optimization techniques in Convolutional Neural Networks». En: *Decision Analytics Journal* 11 (2024), pág. 100470. ISSN: 2772-6622. DOI: https://doi.org/10.1016/j.dajour.2024.100470. URL: https://www.sciencedirect.com/science/article/pii/S2772662224000742.

[58] J. et al. Smith. «Improving classification accuracy through hyperparameter optimization». En: *Heliyon* 10.5 (2024), e26586. DOI: 10.1016/j.heliyon.2024.e26586. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10920154/.

[59] Adam Wiemerslage, Kyle Gorman y Katharina von der Wense. «Quantifying the Hyperparameter Sensitivity of Neural Networks for Character-level Sequence-to-Sequence Tasks». En: ed. por Yvette Graham y booktitle = Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) Purver Matthew. St. Julian's, Malta: Association for Computational Linguistics, mar. de 2024, "674-689". URL: https://aclanthology.org/2024.eacl-long.40/.

[60] Hwanjun Song et al. «Prestopping: How Does Early Stopping Help Generalization Against Label Noise?» En: *Proceedings of the 37th International Conference on Machine Learning* (2020), págs. 9013-9023. URL: http://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-020.pdf.

[61] J. P. Royston. «An Extension of Shapiro and Wilk's W Test for Normality to Large Samples». En: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31.2 (1982), págs. 115-124. ISSN: 00359254, 14679876. URL: http://www.jstor.org/stable/2347973 (visitado 17-06-2025).

[62] Samuel Sanford Shapiro y Martin B Wilk. «An analysis of variance test for normality (complete samples)». En: *Biometrika* 52.3-4 (1965), págs. 591-611.

[63] SciPy Community. *scipy.stats.shapiro - SciPy Documentation*. 2024. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html.

[64] Lars St, Svante Wold et al. «Analysis of variance (ANOVA)». En: *Chemometrics and intelligent laboratory systems* 6.4 (1989), págs. 259-272.

[65] John W. Tukey. «Comparing Individual Means in the Analysis of Variance». En: *Biometrics* 5.2 (1949), págs. 99-114. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/3001913 (visitado 18-06-2025).

[66] William H. Kruskal y W. Allen Wallis and. «Use of Ranks in One-Criterion Variance Analysis». En: *Journal of the American Statistical Association* 47.260 (1952), págs. 583-621. DOI: 10.1080/01621459.1952.10483441. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441.

[67] SciPy Community. *scipy.stats.kruskal - SciPy Documentation*. 2024. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html#scipy.stats.kruskal.

[68] Henry B Mann y Donald R Whitney. «On a test of whether one of two random variables is stochastically larger than the other». En: *The annals of mathematical statistics* (1947), págs. 50-60.

[69] SciPy Community. *scipy.stats.mannwhitneyu - SciPy Documentation*. 2024. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu.

[70] G Rupert Jr et al. «Simultaneous statistical inference». En: (2012).

[71] Wikipedia. *Corrección de Bonferroni — Wikipedia, La enciclopedia libre*. 2023. URL: https://es.wikipedia.org/w/index.php?title=Correcci%C3%B3n_de_Bonferroni&oldid=155958874.

[72] Hieu-Thi Luong y Junichi Yamagishi. *NAUTILUS: a Versatile Voice Cloning System*. 2020. arXiv: 2005.11004 [eess.AS]. URL: https://arxiv.org/abs/2005.11004.

[73] Tuan Duy Nguyen Le, Kah Kuan Teh y Huy Dat Tran. «Continuous Learning of Transformer-based Audio Deepfake Detection». En: *arXiv e-prints*, arXiv:2409.05924 (sep. de 2024), arXiv:2409.05924. DOI: 10.48550/arXiv.2409.05924. arXiv: 2409.05924 [cs.SD].

[74] Muhammad Umar Farooq y Thomas Hain. *Investigating the Impact of Cross-lingual Acoustic-Phonetic Similarities on Multilingual Speech Recognition*. 2022. arXiv: 2207.03390 [cs.CL]. URL: https://arxiv.org/abs/2207.03390.

[75] *Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)*. es. Diario Oficial de la Unión Europea, L 119, 4 de mayo de 2016. 2016. URL: https://www.boe.es/buscar/doc.php?id=DOUE-L-2016-80807.

Índice de figuras

1.1. Diagrama Esquemático del sintetizador Voder obtenida de [1] 3
1.2. Diagrama de arquitectura del sintetizador TACOTRON2 de [2] 4

2.1. Fotografía de la máquina EUPHONIA[10] 9
2.2. Circuito esquemático del Voder[12] 11

3.1. Los segmentos negros indican segmentos de habla, los segmentos rojos ilustran micropausas (pausas de menos de 0.5 segundos y de al menos 0.1 segundos), y los segmentos amarillos indican macropausas (pausas de 0.5 segundos o más).[27] . . . 23

4.1. Ejemplo de dos espectrogramas. EL primer una voz real, el segundo un modelo entrenado para decir la misma frase[38] 29

5.1. Gráfico que muestra la localización de respiración y los valores de RMS y ZCR[46] 38
5.2. Gráfrico que compara una serie de redes neuronales CNN[50] 41

6.1. Distribución de Precision por combinación de características y arquitectura (1 % datos). 50
6.2. Distribución de Precision por combinación de características y arquitectura (5 % datos). 50
6.3. Distribución de Precision por combinación de características y arquitectura (10 % datos). 50
6.4. Top3 conjuntos de características para Precision 51
6.5. Distribución de Recall por combinación de características y arquitectura (1 % datos). 53
6.6. Distribución de Recall por combinación de características y arquitectura (5 % datos). 53
6.7. Distribución de Recall por combinación de características y arquitectura (10 % datos). 53
6.8. Top3 conjuntos de características para Recall 54
6.9. Distribución de Specificity por combinación de características y arquitectura (1 % datos). 56
6.10. Distribución de Specificity por combinación de características y arquitectura (5 % datos). 56

6.11. Distribución de Specificity por combinación de características y arquitectura (10 % datos). 56
6.12. Top3 conjuntos de características para Specificity 57
6.13. Distribución de Accuracy por combinación de características y arquitectura (1 % datos). 59
6.14. Distribución de Accuracy por combinación de características y arquitectura (5 % datos). 59
6.15. Distribución de Accuracy por combinación de características y arquitectura (10 % datos). 59
6.16. Top3 conjuntos de características para Accuracy 60
6.17. Distribución de F1-Score por combinación de características y arquitectura (1 % datos). 62
6.18. Distribución de F1-Score por combinación de características y arquitectura (5 % datos). 62
6.19. Distribución de F1-Score por combinación de características y arquitectura (10 % datos). 62
6.20. Top3 conjuntos de características para F1Score 63
6.21. Distribución de AUC-ROC por combinación de características y arquitectura (1 % datos). 65
6.22. Distribución de AUC-ROC por combinación de características y arquitectura (5 % datos). 65
6.23. Distribución de AUC-ROC por combinación de características y arquitectura (10 % datos). 65
6.24. Top3 conjuntos de características para AUC-ROC 66

7.1. Esquema de NAUTILUS[72] 68