



EDUCACIÓN

SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO
NACIONAL DE MÉXICO

Tecnológico Nacional de México

**Centro Nacional de Investigación
y Desarrollo Tecnológico**

Tesis de Maestría

Modelo de diferenciación de voz humana y sintética

presentada por

ING. Jibran Zaedt Rodriguez Garcia

como requisito para la obtención del grado de
Maestría en Ciencias de la Computación

Director de tesis

Dra. Andrea Magadán Salazar

Codirector de tesis

Dr. Noé Alejandro Castro Sánchez

Cuernavaca, Morelos, México. Febrero de 2026.



Centro Nacional de Investigación y Desarrollo Tecnológico
Departamento de Ciencias Computacionales

Cuernavaca, Mor., 29/Enero/2026

OFICIO No. DCC/013/2026

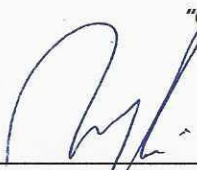
Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de **Jibrán Zaedt Rodríguez García** con número de control **M23CE077**, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"Modelo de diferenciación de voz humana y sintética"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

ATENTAMENTE

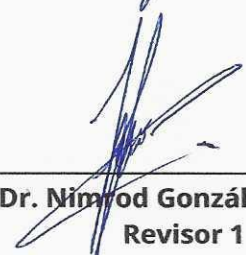
Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"



Dra. Andrea Magadán Salazar
Directora de tesis



Dr. Noé Alejandro Castro Sánchez
Codirector de tesis



Dr. Nimrod González Franco
Revisor 1



Dr. Jonathan Villanueva Távira
Revisor 2

C.c.p. Depto. Servicios Escolares.
Expediente / Estudiante



2026
año de
**Margarita
Maza**



Centro Nacional de Investigación y Desarrollo Tecnológico
Subdirección Académica

Cuernavaca Mor, 29/enero/2026

Oficio No. SAC/035/2026

Asunto: Autorización de impresión de tesis

JIBRÁN ZAEDT RODRÍGUEZ GARCÍA
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS
DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Modelo de diferenciación de voz humana y sintética"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

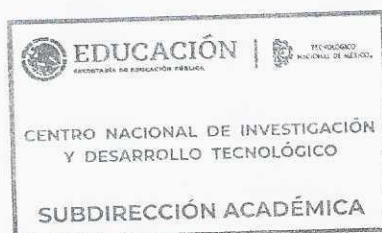
Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

ATENTAMENTE

Excelencia en Educación Tecnológica®

"Conocimiento y Tecnología al servicio de México"

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO



c.c.p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz



2026
año de
Margarita
Maza

cenidet
Centro Nacional de Investigación
y Desarrollo Tecnológico



DEDICATORIAS

A mis padres, quienes han sido el pilar más fuerte en cada paso que he dado. Gracias por su amor incondicional, por sus consejos en los momentos de duda y por enseñarme con su ejemplo que el esfuerzo y la constancia son el camino para alcanzar los sueños. Este logro también es de ustedes, porque han estado presentes en cada madrugada de desvelo, en cada palabra de aliento y en cada abrazo que me devolvió la fuerza cuando más lo necesitaba.

A mis compañeros, con quienes compartí largas jornadas de estudio, dudas, desvelos y también risas. Gracias por demostrarme que el aprendizaje compartido siempre es más enriquecedor, por su apoyo cuando las ideas parecían no avanzar y por esa complicidad que convirtió cada reto en una oportunidad de crecer juntos.

A mis amigos, quienes supieron cuándo distraerme con una charla, una broma o un respiro necesario. Gracias por recordarme que también es importante pausar, sonreír y disfrutar del camino. Su compañía me permitió encontrar equilibrio entre la exigencia académica y los momentos de descanso.

Dedico también este trabajo a todas las personas que, de una u otra manera, me motivaron a seguir adelante, incluso sin darse cuenta. A quienes creyeron en mí cuando yo mismo dudaba, y a quienes me inspiraron a soñar en grande y no conformarme con menos de lo que sé que puedo alcanzar.

Finalmente, me lo dedico a mí mismo, porque detrás de cada página hay horas de esfuerzo, noches de desvelo y la certeza de que la dedicación y la pasión siempre dejan huella.

AGRADECIMIENTOS

A Dios por permitir que esté cumpliendo con un logro más en mi vida y por estar juntos en familia en estos tiempos difíciles de pandemia, gracias Dios!

Al Tecnológico Nacional de México, campus Cenidet por haberme brindado la oportunidad de continuar con mis estudios y superación académica.

A la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) por la oportunidad de formar parte del programa de becas para estudiar la Maestría en Ciencias de la Computación.

A cada uno de mis profesores del Cenidet por su tiempo, asesorías, apoyo y por compartir sus conocimientos.

A mi comité: el Dr. Jonathan Villanueva Tavira y el Dr. Nimrod González Franco por permitir que este trabajo de tesis culmine y por todos sus comentarios, observaciones que me ayudarán a ir creciendo poco a poco y poder llegar a ser un profesional como ustedes.

Agradecer especialmente a la Dra. Andrea Magadán Salazar por haber depositado su confianza desde aquella entrevista que tuvimos, quien finalmente se convirtió en mi guía durante estos dos años. Agradezco mucho su atención, tiempo, dedicación y trabajo en equipo, gracias a usted estoy aquí cumpliendo un logro más.

RESUMEN

El presente trabajo se enfoca en el análisis y comparación de distintas técnicas para el reconocimiento de voces humanas y voces generadas sintéticamente mediante herramientas de clonación. Aunque el objetivo inicial fue desarrollar un sistema de reconocimiento, el proyecto derivó en una investigación experimental centrada en la caracterización de señales de voz, específicamente mediante la estimación de la frecuencia fundamental híbrida (hF0), como descriptor para diferenciar entre voces naturales y clonadas.

Debido a la ausencia de un conjunto de datos confiable y equilibrado, se generó un *dataset* propio a partir de cuatro herramientas de clonación de voz ampliamente utilizadas: RVC, Voice.ai, ElevenLabs y Applio. Se recolectaron un total de 1,256 muestras de voz, equilibradas por idioma (español e inglés) y género (masculino y femenino).

Los resultados experimentales muestran que la frecuencia fundamental constituye una característica relevante para la detección de voces clonadas, siendo su efectividad dependiente de la forma en que es representada y procesada. En particular, la representación de la hF0 mediante histogramas permitió obtener los mejores resultados de clasificación, alcanzando una precisión del 95%, mientras que otras variantes basadas en estadísticas simples o en salidas directas del modelo CREPE presentaron desempeños moderados. Estos hallazgos confirman el potencial del descriptor hF0 para la identificación de clonación de voz y resaltan la importancia de una adecuada extracción y representación de características acústicas. Asimismo, la construcción de un *dataset* diversificado con múltiples herramientas permitió reducir el sesgo asociado al uso de una sola técnica de clonación, proporcionando una base más robusta para futuros estudios en el área de la detección de *deepfakes* de voz.

ABSTRACT

This work focuses on the analysis and comparison of different techniques for the recognition of human voices and synthetically generated voices produced using voice cloning tools. Although the initial objective was to develop a recognition system, the project evolved into an experimental study centered on the characterization of voice signals, specifically through the estimation of the hybrid fundamental frequency (hF0) as a descriptor to differentiate between natural and cloned voices.

Due to the lack of a reliable and balanced dataset, a proprietary dataset was created using four widely used voice cloning tools: RVC, Voice.ia, ElevenLabs, and Applio. A total of 1,256 voice samples were collected, balanced by language (Spanish and English) and gender (male and female).

Experimental results show that the fundamental frequency constitutes a relevant feature for the detection of cloned voices, with its effectiveness depending on how it is represented and processed. In particular, representing hF0 through histograms yielded the best classification results, achieving an accuracy of 95%, while other variants based on simple statistical features or direct outputs from the CREPE model showed moderate performance. These findings confirm the potential of the hF0 descriptor for voice cloning identification and highlight the importance of proper extraction and representation of acoustic features. Additionally, the construction of a diversified dataset using multiple cloning tools helped reduce the bias associated with relying on a single cloning technique, providing a more robust foundation for future studies in the field of voice deepfake detection.

CAPÍTULO 1 INTRODUCCIÓN

En este capítulo se exponen: el análisis del problema, los objetivos (tanto general y como específicos), los alcances, las limitaciones, la selección de herramientas, la metodología de solución propuesta y una descripción breve del resto de capítulos que integran el presente trabajo de tesis.

1.1 Introducción

Los avances en la inteligencia artificial (IA) han transformado múltiples áreas, entre ellas la generación artificial de voz. En particular, las herramientas de clonación de voz han evolucionado a tal punto que permiten crear audios altamente realistas, lo que facilita la elaboración de contenido falso desde prácticamente cualquier lugar. Este fenómeno plantea nuevos desafíos, ya que, a diferencia de las imágenes o los videos, en los cuales suelen detectarse anomalías visuales, los audios falsificados pueden pasar desapercibidos incluso para oyentes atentos.

El análisis específico de la voz como único canal de detección ha sido poco explorado. Ejemplo de ello es la competencia organizada por Meta [1], centrada principalmente en la detección de *deepfakes* en video, cuyos resultados en la identificación de audio falsificado fueron limitados. Esto evidencia una importante área de oportunidad en el desarrollo de tecnologías orientadas exclusivamente al análisis de la señal de voz.

El impacto de los audios sintéticos generados por IA ya se ha manifestado en contextos sociales y políticos. Las redes sociales enfrentan grandes dificultades para moderar este tipo de contenido, y los verificadores humanos suelen ser incapaces de distinguir entre una voz real y una clonada. Un caso representativo ocurrió durante las elecciones parlamentarias en Eslovaquia, donde la organización Reset [2], dedicada al estudio del impacto de las redes sociales en la democracia, identificó la circulación de audios falsificados en los días previos a la votación [3], [4]. En dichas grabaciones, difundidas en plataformas como Facebook, Instagram, Telegram y TikTok, se manipuló la voz del líder político Michal Šimečka [5], logrando más de 50,000 visualizaciones según la biblioteca de anuncios de Facebook [6].

Frente a este panorama, se vuelve imprescindible contar con métodos y herramientas que permitan detectar si un audio ha sido generado mediante inteligencia artificial. Esta tesis se centra precisamente en el análisis y comparación de técnicas que buscan identificar voces sintéticas a partir de sus características acústicas, con especial énfasis en el comportamiento de la estimación de frecuencia fundamental híbrida (hF0) como descriptor distintivo para reconocer clonaciones de voz.

1.2 Análisis del problema

En los últimos años, el desarrollo de tecnologías de inteligencia artificial ha permitido la generación de voces sintéticas casi indistinguibles de una voz humana real. Sin embargo,

a diferencia de otras áreas como la generación de imágenes o videos manipulados, en las cuales existen señales visuales detectables como artefactos, incoherencias anatómicas o patrones de ruido, en el caso del audio, particularmente la voz, aún no se cuenta con una documentación suficiente ni con herramientas que permitan identificar de manera precisa qué características específicas pueden delatar una manipulación.

Actualmente, existen múltiples modelos de generación de voz basados en redes neuronales profundas, como los sistemas de clonación de voz o generación de texto a voz realista. Estas herramientas, aunque útiles para múltiples fines, también representan un riesgo en contextos donde la autenticidad de la voz es crítica, como en la seguridad, el periodismo, la política o la justicia. A pesar de esto, se ha dado poca atención al análisis detallado de los rasgos acústicos que podrían distinguir una voz generada artificialmente de una voz humana.

Esto plantea una necesidad urgente de estudiar las características técnicas y estadísticas que presenta una voz manipulada en comparación con una voz real. De manera particular, surge la interrogante sobre qué tan confiables pueden ser ciertas métricas acústicas (como el *jitter*, *shimmer* o la desviación estándar de la frecuencia fundamental) para detectar voces generadas por modelos de inteligencia artificial, y si estas variables podrían aportar elementos suficientes para clasificar correctamente una voz como auténtica o manipulada.

Ante esta falta de documentación y análisis específico, esta investigación busca explorar cuáles son las características de la voz que permiten detectar algún tipo de manipulación mediante herramientas automáticas de clasificación, con el objetivo de aportar evidencia empírica al campo del análisis de voces sintéticas.

1.3 Complejidad del problema

El desarrollo del sistema propuesto presenta diversos desafíos, debido a que se deben considerar múltiples factores que influyen en la detección de voces clonadas y su diferenciación respecto a voces humanas. Entre los aspectos más relevantes se encuentran:

- Analizar audios en diferentes idiomas, específicamente español e inglés, lo cual implica reconocer variaciones fonéticas, acentos y estructuras lingüísticas propias de cada lengua.
- Considerar voces de distintos sexos, ya que las características acústicas como la frecuencia fundamental, la entonación y el timbre varían significativamente entre voces masculinas y femeninas, lo cual puede afectar la precisión del sistema de detección.
- Analizar audios generados por distintas herramientas de clonación, las cuales emplean modelos de síntesis con niveles variables de realismo, calidad y estilo de generación, dificultando la identificación de patrones comunes.
- Trabajar con una limitada disponibilidad de *datasets* públicos de voces clonadas, lo que obliga a construir o recopilar manualmente conjuntos de audios diversificados para entrenamiento y validación del sistema.

- Falta de documentación técnica sobre las herramientas de clonación de voz y las características específicas que presentan los audios sintéticos, lo que incrementa la dificultad de establecer criterios objetivos para distinguir entre voces reales y generadas.

1.4 Objetivos

1.4.1 Objetivo general

Desarrollar un sistema que realice el reconocimiento de voz natural y de voz generada de forma sintética en al menos 2 lenguajes (español e inglés).

1.4.2 Objetivos específicos

- Estudiar en el estado del arte los diferentes enfoques utilizados en la caracterización de la voz, mediante el análisis de señales y con aprendizaje profundo.
- Reportar el análisis de herramientas de Inteligencia Artificial que permiten sintetizar *deepfakes* de voces.
- Contar con conjuntos de audios de voces creadas por herramientas de Inteligencia Artificial para la clonación de voz y de audios de personas.
- Analizar técnicas de extracción de características de los audios para determinar si la voz procede de un sintetizador o un humano.
- Implementar un algoritmo de clasificación para el reconocimiento de la voz natural y sintética.
- Evaluar el rendimiento del sistema desarrollado de manera cuantitativa a través de las métricas tradicionales en el reconocimiento de patrones.

1.5 Alcances

- El sistema trabaja con audios en un ambiente no controlado.
- El sistema sólo realiza el reconocimiento de voz natural y de voz sintética en los lenguajes español e inglés.
- El sistema debe analizar y evaluar los audios creados de al menos 2 herramientas gratuitas de generación de voz.
- El objetivo es desarrollar un sistema preciso y eficiente en términos de recursos computacionales, permitiendo su potencial portabilidad a dispositivos móviles u otros entornos con recursos limitados.

1.6 Limitaciones

- El audio solo debe contener la voz de un hablante y sin cambios de lenguaje durante la grabación.
- El diseño del clasificador está sujeto a las características del equipo computacional disponible por el autor.
- El sistema no interpreta las palabras presentes en el audio.
- La voz de la persona tiene que ser audible.
- El sistema desarrollado no proporciona información relacionada a la identidad de la persona, género, edad y emoción presente.
- Se consideran los audios de personas dentro del rango de edad de 18 hasta los 60 años.

1.7 Justificación y beneficios

En los últimos años, el uso de tecnologías de clonación de voz ha aumentado considerablemente, impulsado por avances en inteligencia artificial y aprendizaje profundo. Si bien estas herramientas ofrecen beneficios en ámbitos como la accesibilidad, el entretenimiento y la educación, también han abierto la puerta a nuevos riesgos asociados con la suplantación de identidad, el fraude y la manipulación de información mediante el uso de voces sintéticas altamente realistas.

A diferencia de otras áreas como la detección de imágenes falsas, donde ya existen múltiples estudios y herramientas desarrolladas, el análisis detallado de las características acústicas que podrían permitir diferenciar una voz humana de una clonada sigue siendo un campo poco explorado. Esto se ve agravado por la escasa disponibilidad de documentación técnica sobre las herramientas de clonación y la falta de *datasets* públicos que permitan estudiar y evaluar dichas voces.

Por ello, esta investigación se justifica al proponer el desarrollo de un sistema capaz de reconocer voces naturales y sintéticas, analizando audios en diferentes idiomas y generados por diversas herramientas de clonación. Este enfoque permite no solo avanzar en el entendimiento de las particularidades acústicas de la voz clonada, sino también sentar las bases para la creación de herramientas automatizadas de detección, que puedan aplicarse en áreas como la seguridad digital, la verificación biométrica o el periodismo.

Entre los principales beneficios de este trabajo se encuentran:

- Contribuir al conocimiento científico en el área del análisis de voz y su manipulación por IA.
- Ofrecer una herramienta experimental que permita identificar voces sintéticas con base en características acústicas reales.
- Incentivar la construcción de *datasets* diversificados que permitan futuras investigaciones en esta línea.
- Aumentar la seguridad en aplicaciones sensibles, como sistemas de verificación por voz o comunicaciones oficiales.
- Fomentar el uso responsable de las tecnologías de clonación de voz, al visibilizar sus riesgos y potenciales abusos.

1.8 Herramientas seleccionadas

Se realizó una revisión de distintas herramientas orientadas a la generación de audios de voz clonada, con el objetivo de identificar aquellas capaces de producir resultados con características acústicas cercanas a la voz humana. Para la identificación de dichas

herramientas, se consultaron fuentes especializadas, incluyendo repositorios de código abierto como GitHub [7] y comunidades técnicas en plataformas como Discord [8].

1.8.1 Criterios de selección

Los criterios considerados fueron los siguientes:

- **Accesibilidad:** Se priorizó la accesibilidad, buscando herramientas que fueran

preferiblemente gratuitas o que ofrecieran pruebas gratuitas para su evaluación inicial.

- Variedad de modificaciones: Se valoró la capacidad de las herramientas para realizar una variedad de modificaciones. Dado que los mismos parámetros no siempre generan resultados óptimos para todas las palabras, era fundamental que las herramientas permitieran ajustar diversos parámetros para lograr el mejor rendimiento en la clonación de voz.
- Calidad de los audios de salida: Se consideró fundamental que los audios generados mantuvieran la máxima naturalidad posible, evitando cualquier tipo de entonación robótica o artificial.

Las herramientas de clonación de voz y los algoritmos para la extracción de tono vocal emplean técnicas para generar audios que imitan voces humanas correspondientes a individuos reales. Aunque comparten ciertas similitudes, es importante destacar que una diferencia fundamental radica en que, una vez se ha creado el modelo de voz de una persona, no es necesario que esta hable para producir audios que la representen.

El análisis de estas herramientas y algoritmos no solo amplía la comprensión de las aplicaciones potenciales de la clonación de voces, sino que también facilita el acceso a recursos para usuarios en todo el mundo, permitiéndoles utilizar estas voces para una variedad de propósitos.

Por otro lado, las herramientas de clonación de voz proporcionan a los usuarios la capacidad de personalizar y generar audios de voz según especificaciones detalladas. Estas herramientas permiten ajustes en aspectos como el tono, el ritmo, el acento y otros atributos vocales. No solo se utilizan para fines tecnológicos, sino que también se integran en proyectos creativos, comerciales y de entretenimiento, ofreciendo una funcionalidad más interactiva y específica según las necesidades del usuario.

1.8.2 Herramientas utilizadas

A continuación, se muestran las herramientas que cumplieron con los criterios para ser utilizadas

1.8.2.1 *Applio* [9]

Esta aplicación es de uso público y gratuito, mediante la cual los usuarios pueden crear modelos de clonación de voz sin límite alguno. Además, ofrece otras herramientas como: la descarga de modelos de voces ya entrenados y listos para ser utilizados, la capacidad de transformar un audio utilizando dichos modelos, la opción de escribir texto y reproducirlo con la voz del modelo seleccionado, y diversos algoritmos de extracción de tono, como se ilustra en la Figura 1.1

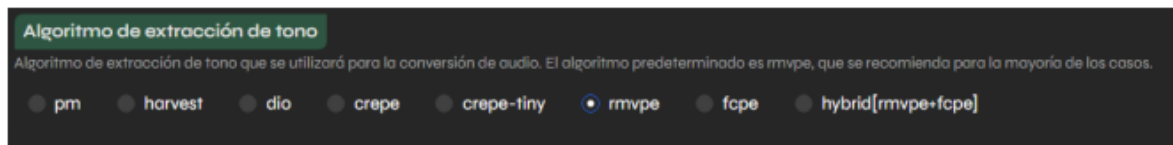


Figura 1.1 Algoritmos de extracción de tono de la herramienta Applio [9]

1.8.2.2 **RVC [10]**

Esta aplicación, que cuenta con una versión web, se encuentra disponible de forma gratuita para el público en general. A través de ella, los usuarios pueden crear modelos de clonación de voz sin restricciones. Además de esta función principal, la aplicación ofrece diversas herramientas complementarias, entre las que se incluyen la transformación de audios utilizando la voz seleccionada, la generación de audio a partir de texto mediante los modelos disponibles y la aplicación de distintos algoritmos para la extracción del tono vocal, tal como se muestra en la Figura 1.2. Asimismo, la tasa de muestreo puede configurarse entre 40 kHz y 48 kHz.

En la versión local del programa se pueden utilizar los siguientes algoritmos:

- **Pm (Pitch Marking):** Algoritmo empleado en el procesamiento del habla para detectar variaciones en la frecuencia fundamental. Resulta útil para el análisis de la calidad vocal y la síntesis de voz. En el sistema, su uso en señales de canto permite un procesamiento más rápido con alta calidad, a costa de un mayor uso de CPU.
- **Harvest:** Método empleado en el procesamiento del habla para la estimación precisa de la frecuencia fundamental y la extracción de características de la señal. En el sistema, proporciona resultados de mayor calidad, a costa de un mayor tiempo de procesamiento.
- **DIO (Distributed Input Output):** Método de estimación de la frecuencia fundamental basado en técnicas de procesamiento paralelo o distribuido en sistemas informáticos. En la aplicación, resulta útil para acelerar el proceso de análisis de la señal.
- **Rmvpe (Robust Model for Vocal Pitch Estimation in Polyphonic Music):** Modelo diseñado para la estimación del tono vocal en música polifónica, con capacidad para extraer de manera robusta la frecuencia fundamental en presencia de distintos tipos de ruido. En la aplicación, ofrece el mejor desempeño en términos de calidad del resultado, además de requerir un menor consumo de recursos de CPU y GPU.
- **Rmvpe_gpu:** Variante del modelo Rmvpe que emplea procesamiento acelerado por GPU, lo que implica un mayor consumo de recursos de CPU y GPU. En la aplicación, permite seleccionar el número de procesos de CPU utilizados para la extracción de tono y el procesamiento de datos, con un rango configurable de 0 a 8.



Figura 1.2 Algoritmos de extracción de tono de la herramienta RVC [10].

1.8.2.3 **Voice.ai [11]**

Esta aplicación cuenta con una amplia variedad de voces desarrolladas por la comunidad. Permite a los usuarios realizar cambios de voz en tiempo real y crear sus propios modelos de voces. Aunque es una aplicación de pago, ofrece la posibilidad de adquirir monedas diariamente al iniciar sesión, las cuales pueden ser utilizadas para acceder a las voces. Una vez que se ha adquirido una voz, esta puede ser utilizada sin restricciones, en la Figura 1.3 se puede observar la interfaz, así como algunas de las voces más utilizadas.

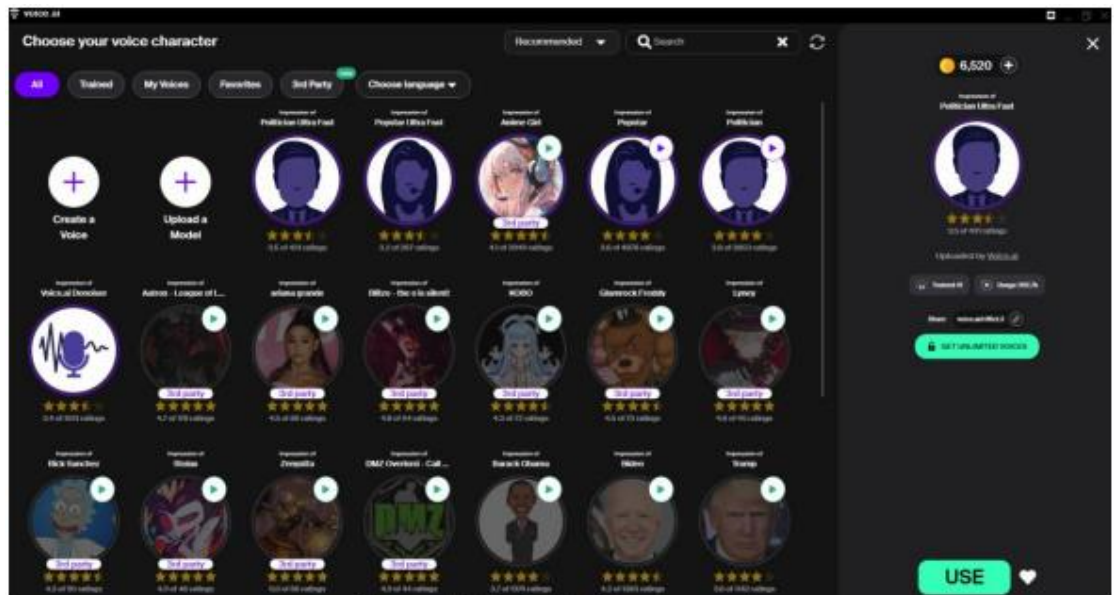


Figura 1.3 Voice.ai interfaz de usuario [11].

1.8.2.4 Elevenlabs [12]

ElevenLabs es una avanzada herramienta de Inteligencia Artificial (IA) que ofrece tecnologías de Texto a Voz, Voz a Voz y Clonación de voz. Con esta aplicación, es posible generar audio hablado de alta calidad en una variedad de voces, estilos e idiomas, así como ajustar géneros, edades, tonos y acentos según las preferencias del usuario. Su modelo de IA captura de manera excepcional la entonación y las inflexiones humanas, ofreciendo una experiencia de voz sumamente realista.

Una de las características destacadas de ElevenLabs es su amplia biblioteca de voces sintéticas compartidas por la comunidad, así como la función VoiceLab, que permite diseñar nuevas voces sintéticas o clonar voces existentes. Además, su API completa y bien documentada facilita su integración en proyectos de desarrollo y automatizaciones.

En resumen, ElevenLabs es una plataforma completa que aprovecha la IA para proporcionar una experiencia de voz de alta calidad, consolidándose como una herramienta líder en el ámbito de la voz.

1.9 Metodología de solución

La metodología general se compone de tres etapas: procesamiento, extracción de características y clasificación. A partir de esta estructura común, se desarrollaron cuatro variantes experimentales, las cuales difieren principalmente en la forma de extraer, representar y analizar las características acústicas de los audios.

1.9.1 Procesamiento

- Todos los audios fueron normalizados y segmentados a una duración estándar de 10 segundos, garantizando uniformidad para el análisis posterior.
- Se verificó que todos los archivos estén en formato .WAV, transformando aquellos que no cumplieran esta condición mediante scripts en Python.
- Para una de las variantes, se incorporó la aplicación de filtros digitales de tipo Butterworth de quinto orden para resaltar componentes relevantes de la señal vocal y reducir el ruido.

1.9.2 Extracción de características

- Se emplearon los siguientes filtros digitales: un filtro pasa baja con atenuación de frecuencias superiores a 500 Hz, un filtro pasa alta con atenuación de frecuencias inferiores a 2000 Hz y un filtro pasa banda que permite el paso de frecuencias entre 500 Hz y 2000 Hz. Dichos filtros se aplicaron después de la normalización temporal y antes del procesamiento destinado a la extracción de características.
- En [13] se señala que existe una diferencia significativa entre la frecuencia fundamental (f_0) promedio de hombres y mujeres. En términos generales, la f_0

promedio es notablemente más alta en mujeres y el rango de f_0 en Hz es más amplio en comparación con el de los hombres. A pesar de que las diferencias en el rango de f_0 entre sexos desaparecen cuando se expresan en semitonos o como un factor de modulación, los valores promedio de f_0 continúan siendo distintos. En resumen, la f_0 promedio de los hombres no es equivalente al f_0 en las mujeres.

- El factor de modulación, en el contexto de la voz, se refiere a la variación de la frecuencia fundamental (f_0) de la voz a lo largo de un período de tiempo. Esta medida indica cuánto fluctúa la frecuencia fundamental alrededor de su valor promedio. La modulación de frecuencia puede ser cuantificada en porcentaje o en unidades de frecuencia (como Hz). Un alto factor de modulación sugiere que la voz presenta numerosas variaciones en tono, lo cual puede interpretarse como expresividad o variabilidad vocal.

En las cuatro variantes de la metodología se utilizó el algoritmo CREPE para la extracción de la frecuencia fundamental (f_0), aplicando distintas estrategias de análisis y considerando diferentes versiones del modelo (A y B) según la implementación utilizada

- En la primera variante, se empleó el algoritmo CREPE A (versión *Tiny*) implementado en MATLAB [14] donde se calcula hF0 (estimación de frecuencia fundamental híbrida) y se utiliza el vector de valores para generar un histograma.
- En la segunda variante, también utilizando CREPE A en MATLAB, se calculó la desviación estándar de la f_0 por segmento, considerando valores con y sin ceros, como una representación estadística de la variabilidad tonal. En ambas variantes basadas en MATLAB, se utilizó la versión *Tiny* del modelo CREPE, dado que es la única disponible en dicha implementación.
- En la tercera variante, se usó CREPE B implementado en Python, generando un vector de características directamente a partir de los valores de f_0 extraídos.
- En la cuarta variante, se utilizaron las versiones de CREPE B implementadas en Python. Se extrajo un vector de características desde la f_0 , el cual fue empleado en una estrategia de clasificación distinta, utilizando el algoritmo XGBoost. En Python se pueden encontrar distintas variantes de CREPE y se evaluaron las versiones *Tiny*, *Small*, *Medium*, *High* y *Full*, lo que permitió analizar el impacto del tamaño del modelo sobre la precisión de la estimación de la frecuencia fundamental.

1.9.3 Clasificación

En las variantes uno, dos y tres, los vectores de características fueron clasificados mediante una Máquina de Vectores de Soporte (SVM). En la cuarta variante, se utilizó el algoritmo XGBoost para la clasificación debido a la gran cantidad de datos y los tiempos de entrenamiento. En todos estos casos, se aplicó una división del

conjunto de datos en un 80% para entrenamiento y 20% para validación.

1.10 Contenido

Este trabajo de tesis está conformado por seis capítulos:

- En el capítulo 1 se presentan los objetivos generales y específicos, la descripción del problema relacionado con la creciente generación de voces sintéticas, así como la metodología de solución propuesta para desarrollar un sistema capaz de diferenciar voz humana y voz clonada mediante técnicas de procesamiento de señales e inteligencia artificial.
- El capítulo 2 aborda la revisión de antecedentes y artículos científicos relacionados con la caracterización de la voz humana, incluyendo investigaciones sobre detección de edad, clasificación por género y diferenciación entre voces naturales y sintéticas. Para ello, se recurre a áreas como el procesamiento digital de audio, la inteligencia artificial y el aprendizaje profundo, con el objetivo de comprender los patrones acústicos relevantes para detectar manipulación o clonación de voz.
- El capítulo 3 describe el marco teórico del proyecto. Se explica la biología de la producción de la voz, abarcando el funcionamiento de las cuerdas vocales, el fuelle respiratorio y los resonadores. Posteriormente, se detalla el proceso de modelación digital de la señal de voz, desde el muestreo hasta la cuantización. También se presentan las métricas de evaluación empleadas para medir el rendimiento de los clasificadores, como exactitud, precisión, *recall* y *F1-score*. Finalmente, se incluyen las librerías, entornos de programación y herramientas utilizadas durante el desarrollo del modelo.
- El capítulo 4 expone el análisis y diseño del sistema planteado, describiendo los elementos presentes en los distintos ambientes de trabajo. Se explican los módulos de preprocesamiento de audio, la creación del conjunto de datos y la preparación necesaria para la experimentación. Asimismo, se detallan las fases que conforman el flujo de entrenamiento y evaluación del modelo.
- El capítulo 5 presenta las configuraciones utilizadas en los experimentos, incluyendo los hiperparámetros y la partición de los conjuntos de entrenamiento y evaluación. Se describen los casos de prueba correspondientes a las variantes de la metodología implementada. Además, se analiza el desempeño de los clasificadores mediante las métricas de evaluación, con el propósito de seleccionar el mejor modelo predictivo y dejarlo preparado para su uso en ambientes reales.
- Finalmente, en el capítulo 6 se exponen las conclusiones generales del proyecto, la justificación de los objetivos alcanzados y los aportes logrados durante la investigación. También se presentan los trabajos futuros derivados del proyecto, así como las actividades académicas adicionales desarrolladas durante el programa de posgrado.

CAPÍTULO 2 ESTADO DEL ARTE

El capítulo aborda la revisión de antecedentes y artículos relacionados con la identificación de características de la voz humana, tales como la detección de edad, la clasificación por género y la diferenciación entre voces naturales y sintéticas.

2.1 Antecedentes

Se realizó la revisión de tesis relacionadas en el repositorio institucional del CENIDET, con respecto al procesamiento de la voz no encontrando trabajos desarrollados. Sin embargo, se presenta la descripción de la tesis en desarrollo titulada “Sistema de reconocimiento de género mediante el análisis acústico de voz” realizada por Enrique Díaz [15].

El objetivo del proyecto es: “Diseñar e implementar un sistema que realice el reconocimiento del género mediante análisis acústico de la voz en al menos 2 lenguajes (español e inglés) adquiridos de un conjunto de datos públicos y grabados en ambientes no controlados”. La metodología se integra de las siguientes etapas:

- Selecciona un conjunto de datos que contiene varios idiomas, para realizar un panorama general de la información de sus elementos (formato de archivo y etiquetas disponibles).
- Construye el conjunto de características acústicas y cepstrales, que permitan inferir el género de una persona y que estén justificadas en la literatura del tema.
- Implementa los métodos de preprocesamiento (*framing*, *windowing* y detector de actividad de voz) para obtener señales digitales (de voz) con una menor cantidad de ruido para poder procesarlas.
- Implementa los métodos de extracción de características.
- Implementa tres clasificadores para el reconocimiento de género en los idiomas: español, inglés, alemán, francés y chino.
- Finalmente, evalúa los resultados mediante las métricas de precisión, sensibilidad, puntaje F1 y área bajo la curva de funcionamiento del receptor.

Como resultados reporta que, una forma de distinguir el género de una voz es a través del dimorfismo sexual; es decir, considerando las variaciones en la fisionomía entre hombres y mujeres. En cuestiones acústicas, la frecuencia de una voz masculina tiene a ser menor que su contraparte femenina. La experimentación presentada muestra que el sistema alcanza reconocimientos superiores al 94% de precisión.

2.2 Análisis del Estado del arte

El reconocimiento de voz ha sido abordado desde distintos enfoques a lo largo del tiempo. De manera general, pueden identificarse cuatro aproximaciones principales: el enfoque acústico-fonético, el enfoque de reconocimiento de patrones, el enfoque basado en inteligencia artificial y el enfoque sustentado en redes neuronales profundas. En los trabajos analizados, estos enfoques se emplean para extraer y modelar características acústicas de la señal de voz con fines de identificación, clasificación o reconocimiento de atributos específicos del hablante. Las Tablas 2.1, 2.2 y 2.3 presentan un resumen comparativo de los estudios revisados, considerando aspectos como el objetivo del trabajo, las características empleadas, los modelos utilizados y los resultados obtenidos. Posteriormente, se presenta una síntesis cualitativa de dichos trabajos.

2.2.1 Sistemas de reconocimiento de voz

La búsqueda se centró en los artículos que hablan sobre cómo detectar deepfake voice, pero debido a los pocos trabajos relacionados se realizó una búsqueda en trabajos de reconocimiento de emociones, género y las arquitecturas utilizadas para la clonación de voz. El objetivo fue identificar diversas características significativas en la voz.

Hay cuatro enfoques principales en el reconocimiento de voz: el enfoque acústico-fonético, el enfoque de reconocimiento de patrones, el enfoque de inteligencia artificial y el enfoque de redes neuronales. Las Tablas 2.1, 2.2 y 2.3 resumen los trabajos analizados y a continuación se encuentran la síntesis de los mismos.

Tabla 2.1 Artículos sobre reconocimiento de deepfakes voice

Artículo	Año	Técnicas	Observaciones	Resultados
Human Perception of Audio Deepfakes [16]	2022	Clasificación humana. Un LSTM bidireccional de tres capas (BiLS). RawN2.	Pese a usar 2 modelos para la clasificación humana solo reportan resultados de un modelo, falta del segundo modelo.	ASVspoof 2019 RawN2 Precisión: 95% Humanos Precisión: 80%
Detecting Deepfake Voice Using Explainable Deep Learning Techniques [17]	2022	CNN CNN-LSTM CNN-LSTM-perm	Se ocupa un modelo de evaluación en el espectrograma para detectar las monotonías de la voz sintética.	ASV multiple VC CNN: 83.61% CNN-LSTM: 84.84% CNN-LSTM-perm: 88.29%
Replay And Synthetic Speech Detection With Res2Net Architect URE [18]	2021	Res2Net50 ResNet34	Se verificó Res2Net50 y ResNet34.	EER=2.502
Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction [19]	2022	Modelo matemático	Con respecto al conjunto de audios con el que se entrenó, solo se menciona como "amplio".	ASVspoof2019 TPR: 92.4%

Tabla 2.2 Artículos sobre reconocimiento de género

Artículo	Año	Características	Clasificador	Resultados
Deep learning of Voice Gender Identification: Proof-of-concept for Gender-Affirming Voice Care [20]	2020	Cepstrales	Red neuronal profunda	Precisión:93% Exhaustividad:95% Puntaje F1: 94 %
Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks [21]	2021	Ceptrales	X-vector with QuartzNet embedder con transferencia de aprendizaje. Y D-vector feature extractor con módulos preentrenados de voces comunes.	Con ambos clasificadores precisión: 99.6%
Performance Analysis of ML Algorithms to Detect Gender Based on Voice [22]	2021	Acústicas	VAD MFCCs with Cepstral Mean Normalization	Mejor precisión: Red neuronal: 83%
A comparative Study of Deep Learning and Machine Learning Approaches in Speech Emotion and Gender Recognition System [23]	2021	Cepstrales	Binary SVMs Multiclass SVM	Inglés Exactitud:97.1%Pr ecisión:97.9 % Exhaustividad: 95.8 % Canarés: Exactitud:84% Precisión:85.4% Exhaustividad: 82%
Voice gender recognition under unconstrained environments using self-attention [24]	2021	Cepstrales	MFCC	Atención pura: Exactitud: 95.11% Precisión: 96.07% Exhaustividad: 96.27% Puntaje F1:96.17% Atención con convolución Exactitud: 96.23% Precisión: 97.06% Exhaustividad: 96.68% Puntaje F1: 97.02%

Tabla 2.3 Artículos sobre reconocimiento de emociones

Artículo	Año	Características	Emociones analizadas	Resultados
A comparative Study of Deep Learning and Machine Learning Approaches in Speech Emotion and Gender Recognition System [23]	2021	• Modulación de los Coeficientes cepstrales defrecuencia mel (MFCC). • Características espectrales. • Tono • Coeficiente cepstral de predicción lineal. • Energía	• Feliz • Triste • Neutral • Disgusto • Enojo • Aburrimiento • Ansiedad	Exactitud en Inglés: 97.24% Canarés:96.77 %
A voice-based real-time Emotion detection Technique using Recurrent neural Network empowered feature modelling [25]	2022	Coeficientes c e p s t r a l e s Frecuencia Mel (MFCC)	• Feliz • Triste • Neutral • Enojado • Emocionado • Frustrado	Precisión: 60,87%
Detection and Analysis Of Emotion from Speech Signals [26]	2015	Coeficientes c e p s t r a l e s Frecuencia Mel (MFCC)	• Neutral • Ira • Alegría • Tristeza	Precisión: entre 37% a 45%.
Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters [27]	2020	Amplitud Tono	• Enojo • Miedo • Felicidad • Neutral • Tristeza • Asco • Sorpresa	RF:95%
Perceptual Cues in Nonverbal Vocal Expressions of Emotion [28]	2010	Amplitud, Tono Perfil espectral de estímulos	• Ira • Miedo • Disgusto • Tristeza • Sorpresa	Precisión: 56.4%

2.2.2 Síntesis de voz: Texto a voz y voz a texto

La síntesis de la voz tomada de texto (TTS) [29] es la simulación del habla humana en una computadora, haciendo referencia a la transformación artificial de texto a un audio, mientras la conversión de voz a texto (STT) permite transformar el lenguaje hablado a texto escrito. Estas tecnologías han sido ampliamente adoptadas en diversos ámbitos, entre los que destacan:

- Servicios de transcripción.
- Asistentes virtuales.
- Herramientas de accesibilidad.

- Servicios de traducción de idiomas.
- Herramientas de dictado.

Los avances tecnológicos han impulsado la disponibilidad de grandes conjuntos de datos, mejorando las técnicas de aprendizaje profundo y una mayor potencia computacional, lo que ayuda en mejoras significativas de precisión, naturalidad y usabilidad de sistemas.

Los avances en esa área crecen, incluso en [30] se menciona que hay competencias para evaluar la naturalidad de la voz, donde se pide a los oyentes que elijan una respuesta que represente qué tan inteligible es la voz sintética en una escala de 1 (completamente ininteligible) a 5 (completamente inteligible).

Se realizó el análisis de algunas herramientas de generación de voz mediante inteligencia artificial, la información se presenta en las Tablas 2.4, 2.5 y 2.6.

Tabla 2.4 Técnicas de síntesis de voz presentadas en [31]

Herramienta	Técnica utilizada	Características	Dataset	Limitaciones
Tacotron [32]	Encoder-Decoder con RNN	Características profundas	Privado (24.6 hrs.)	Costoso entrenar al modelo
Tacotron2 [33]	Encoder-decoder	Características lingüísticas	Japanese speech corpus from the ATR Ximera dataset (46.9 hrs.)	No hace síntesis de voz en tiempo real.
Deep Voice 1 [34]	Red neuronal profunda	Características lingüísticas	Privado (20 hrs.)	El entrenamiento individual de cada módulo provoca un error acumulativo en el discurso sintetizado
Deep Voice 2 [35]	RNN	Características profundas	VCTK (44 hrs.)	Costo de entrenar el modelo
Deep Voice 3 [36]	Encoder-decoder	Características profundas	▪ Privado (20 hrs.) ▪ VCTK (44 hrs.) ▪ LibriSpeech ASR (820 hrs.)	No generaliza correctamente
WaveNet [37]	Red neuronal profunda	▪ características lingüísticas. ▪ frecuencias fundamentales (registro F0)	VCTK (44 hrs.)	Computacionalmente complejo

Tabla 2.5 Herramientas que generan habla a partir de texto.

Nombre	Número de voces gratuitas	Puedes crear modelos	Método de ejecución
app.kits [38]	2	Si	web
Myvocal [39]	1	Si	web
RVC [10]	ilimitado	Si	web/local
APPLIO [9]	ilimitado	Si	local
Speechify [40]	5	No	web
Resemble [41]	10	No	web
CereVoice Me [42]	1	Si	web
Fakeyou [43]	ilimitado	No	web

Tabla 2.6 Herramientas de voz a voz (voice to speech)

Nombre	Número de voces gratuitas	Puedes crear modelos	Puedes modificar características de la voz	método de ejecución
RVC [10]	ilimitadas	Si	• dZb • tono máximo • tono mínimo • filtro de mediana • Postprocesamiento • algoritmo de extracción de tono	web/local
APPLIO [9]	ilimitadas	Si	• dZb • tono máximo • tono mínimo • filtro de mediana • Postprocesamiento • algoritmo de extracción de tono	Local
app.kits [38]	1	Si	No	Web
voice.ai [11]	Ilimitadas por tokens diarios	Si	No	Web
Resemble [41]	0	Si	No	Web
Fakeyou [43]	ilimitado	No	No	Web

2.2.3 Estudios relacionados con la caracterización de la voz

A continuación, se presenta la síntesis de artículos relacionados con el estudio de la voz humana, su caracterización y clasificación. Aunque no abordan directamente la detección de clonación de voz mediante características acústicas específicas, ofrecen un panorama general sobre líneas de investigación vinculadas, como la identificación de edad, género y autenticidad vocal.

En cada trabajo se destacan el objetivo, el enfoque metodológico, los modelos utilizados y los resultados obtenidos, con el fin de comprender el estado actual del área y detectar vacíos que esta tesis busca explorar.

Deep learning of Voice Gender Identification: Proof-of-concept for GenderAffirming Voice Care (Aprendizaje profundo de la identificación de género por voz: prueba de concepto para el cuidado de la voz que afirma el género) [20]

- **Contexto:** La necesidad de una atención de la voz que afirme el género ha aumentado en la población transgénero en la última década. Actualmente, faltan mediciones objetivas de los resultados del tratamiento para evaluar el éxito de estas intervenciones.
- **Metodología:** Se utilizaron doscientas setenta y ocho voces de hablantes masculinos y femeninos de la base de datos Perceptual Voice Qualities para entrenar una red neuronal profunda para clasificar las voces como masculinas o femeninas. Cada muestra de audio se asignó al dominio de frecuencia utilizando espectrogramas Mel.
- **Resultados:** Se obtuvo una precisión general del 92%, tanto al considerar la precisión por espectro como por métrica de paciente. La precisión del modelo fue mayor para reconocer voces femeninas (puntuación F1 de 0,94) en comparación con voces masculinas (puntuación F1 de 0,87).

Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks (Métodos de estimación de género y edad basados en el habla utilizando redes neuronales profundas) [21]

- **Contexto:** El propósito de este artículo es explorar cómo la extracción de información del habla puede ser útil en una variedad de aplicaciones, desde la identificación personal hasta los robots de voz y los asistentes inteligentes. También se discuten los enfoques de aprendizaje profundo utilizados en estos sistemas y cómo pueden mejorar la experiencia del usuario y generar ingresos para las empresas que los utilizan.

- **Metodología:** La Figura 2.1 representa los tres enfoques diferentes, haciendo uso de módulos de clasificación/regresión.

El primer sistema implementado corresponde a una adaptación de la solución del vector X para la predicción conjunta de edad y género. Esta adaptación incluye la sustitución de capas del modelo original y la incorporación de un detector de actividad de voz (VAD). En la etapa de preprocesamiento de la forma de onda se realizó un recorte aleatorio de 5 segundos de duración. Como funciones de entrada se emplearon coeficientes cepstrales en las frecuencias de Mel (MFCC) de 23 dimensiones, aplicando normalización media cepstral sobre ventanas deslizantes de 3 segundos.

El segundo sistema corresponde a la propuesta del autor y, a diferencia del primero, utiliza características MFCC de 30 dimensiones en lugar de 23. Asimismo, en lugar de la normalización media cepstral —la cual se ha demostrado experimentalmente que presenta un desempeño deficiente para expresiones cortas en el conjunto de datos TIMIT—, se aplicó una normalización de decibels relativos a escala completa, fijada en un nivel de -30 dB, directamente sobre las formas de onda recortadas antes de la extracción de características.

La tercera opción utiliza el vector d y, en lugar de una red convolucional profunda, emplea una red neuronal recurrente LSTM multicapa. La salida de la última unidad oculta se utiliza como incrustación del modelo.

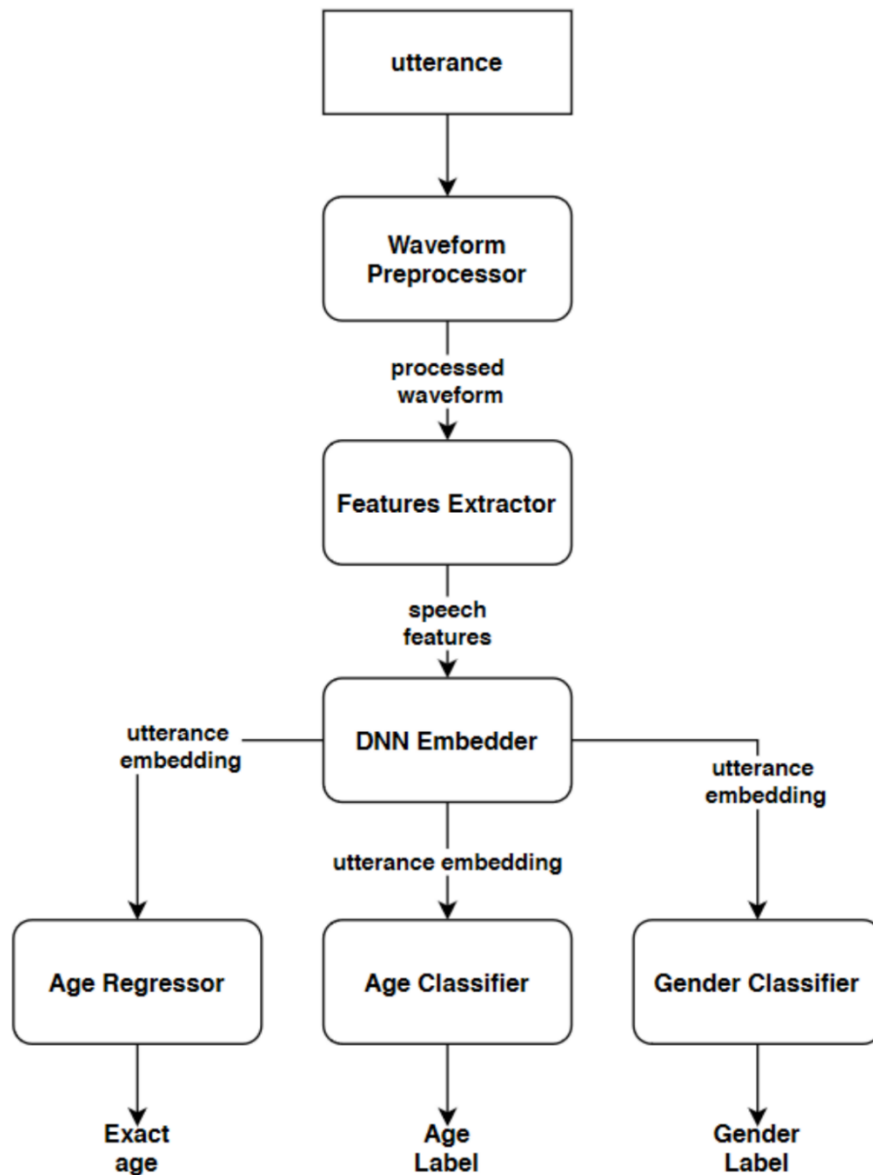


Figura 2.1 Representación del sistema de estimación conjunta de edad y clasificación de género [21].

• **Resultados:** El sistema logró un rendimiento sólidamente alto con precisiones que varían del 96,8% al 99,6%. dependiendo de los conjuntos de datos de entrenamiento y prueba. El resultado más alto se logró cuando el conjunto de datos de Common Voice se utilizó para el entrenamiento, el algoritmo se ajustó aún más en conjunto de datos TIMIT, lo que permitió la precisión de la clasificación al nivel del 99,6% por género reconocimiento.

Performance Analysis of ML Algorithms to Detect Gender Based on Voice (Análisis de rendimiento de algoritmos de aprendizaje de máquina para detectar el género basados en voz) [22]

- **Contexto:** El problema de la clasificación de género fue abordado mediante el análisis de imágenes. A partir de estas, se obtenían múltiples datos que ayudaban a los algoritmos a inferir el género de la persona. Hoy en día, se realiza ese proceso a partir de la señal de voz.

- **Metodología:** 3000 archivos de audio provenientes de Voxforge fueron procesados en este trabajo. Se dividieron en dos conjuntos, uno de entrenamiento (70% del total) y otro de prueba (30% restante). A cada audio se hizo un encuadre de 200 milisegundos y se calcularon sus frecuencias mediante la transformada de Fourier de tiempo reducido en tiempo discreto.

Posteriormente se extrajeron las siguientes características en cada archivo.

- Desviación estándar de la frecuencia.
- Moda de la frecuencia.
- Mediana de la frecuencia.
- Primer cuantil.
- Tercer cuantil.
- Curtosis.
- Centroide espectral.
- Entropía espectral.
- Promedio de la frecuencia fundamental medida a través de la señal acústica.
- Frecuencia fundamental mínima medida a través de la señal acústica.
- Frecuencia fundamental máxima medida a través de la señal acústica.
- Mínimo de frecuencia dominante medida a través de la señal acústica.
- Promedio de la frecuencia dominante medida a través de la señal acústica.
- Máximo de frecuencia dominante medida a través de la señal acústica.
- Máximo de frecuencia dominante medida a través de la señal acústica.
- Modulación.

- **Resultados:** Los algoritmos de árboles de decisión, bosque aleatorio y potenciación del gradiente coincidieron que el primer cuantil es la característica que aporta más información en la clasificación. Los resultados generales pueden verse en la Tabla 2.7.

Tabla 2.7 Resultados de los clasificadores [22].

Algoritmos	Exactitud (accuracy)	Precisión (precision)	Exhaustividad (recall)	Puntuación F1 (F1 score)
Bosque aleatorio	89 %	81 %	95 %	88 %
Árbol de decisión	82 %	77 %	79 %	64 %
Máquina de soporte vectorial	88 %	82 %	87 %	75 %
Red neuronal	89 %	83 %	90 %	86 %
Potenciación del gradiente	90 %	82 %	95 %	88 %

A comparative Study of Deep Learning and Machine Learning Approaches in Speech Emotion and Gender Recognition System (Un estudio comparativo de los enfoques de aprendizaje profundo y aprendizaje automático en el sistema de reconocimiento de género y emoción del habla) [23]

- **Contexto:** El reconocimiento de emociones y género mediante el habla se enfrenta a diversos retos. Uno de estos es cómo seleccionar las características adecuadas y que permitan distinguir el género y la emoción humana ante múltiples escenarios.
- **Metodología:** Se construyeron dos conjuntos de datos. El primero de hablantes del inglés (192 hombres y 214 mujeres) y el segundo de canarés (136 hombres y 146 mujeres). Para cada archivo de audio se extrajeron 39 características de los coeficientes cepstrales de frecuencias de Mel y extraídas por la librería librossa de python.
- **Resultados:** Los resultados del sistema clasificador de género se pueden ver en la Tabla 2.8 y para la clasificación de emociones en la Tabla 2.9.

Tabla 2.8 Resultados de clasificación de género usando el modelo de mezclas gaussianas [23]

	Precisión	Exhaustividad	Exactitud
Inglés	97.9%	95.8%	97.1%
Canarés	85.4%	82%	84%

Tabla 2.9 Resultados de la clasificación de emociones con la métrica: Exactitud [23]

Inglés				
Aprendizaje de maquina				
Arboles de decisión 46%	Bosque aleatorio 57%	Bayesiano ingenuo 37%	Máquina de soporte vectorial 48%	K vecinos más cercanos 53%
Aprendizaje profundo				
Perceptrón multicapa 62%	Convolutacional de una dimensión 70%		Convolutacional de dos dimensiones 97.24%	
Canarés				
Aprendizaje de maquina				
Arboles de decisión 45%	Bosque aleatorio 55%	Bayesiano ingenuo 33%	Máquina de soporte vectorial 45%	K vecinos más cercanos 51%
Aprendizaje profundo				
Perceptrón multicapa 61.84%	Convolutacional de una dimensión 70.51%		Convolutacional de dos dimensiones 96.77%	

Voice gender recognition under unconstrained environments using selfattention (Reconocimiento de género de voz en entornos sin restricciones utilizando la atención propia) [24]

• **Contexto:** El reconocimiento de género por voz es una tarea no trivial que ha sido estudiada ampliamente. No obstante, cuando la voz está rodeada de múltiples ruidos, la tarea se vuelve más compleja.

• **Metodología:** Los audios fueron obtenidos del conjunto de datos VoxCeleb. Se construyeron dos subconjuntos: el primero de 12000 audios (6538 hombres y 5462 mujeres) y el segundo de 25000 audios (13289 hombres y 11711 mujeres). Posteriormente cada conjunto se subdividió en dos conjuntos. Uno de entrenamiento y otro de validación.

Cada audio se ajustó para que su duración fuera de 5 segundos. Posteriormente se les extrajeron 40 coeficientes de frecuencias de Mel en intervalos de 31 milisegundos. Posteriormente fueron visualizados en un espectrograma.

Se diseñaron dos modelos para este trabajo.

- El primero consistió en 6 capas de atención propia seguida de una capa densa de 2048 unidades. Como clasificador se utilizó una regresión logística.
- El segundo modelo consistió en la modificación del primer modelo agregando al inicio de estos dos módulos, uno de reducción y el otro de residuales de inicio.

- **Resultados:** Los resultados de ambos modelos se muestran en la Tabla 2.10.

Tabla 2.10 Resultados de los modelos de atención pura y convolución [24]

	Exactitud	Precisión	Exhaustividad	Puntaje F1	Área bajo ROC
Atención pura	95.11%	96.07%	96.27%	96.17%	98.36%
Convolución con atención propia	96.23%	97.06%	96.68%	97.02%	98.80%

A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling (Una técnica de detección de emociones en tiempo real basada en voz que utiliza modelado de características potenciado por redes neuronales recurrentes) [25]

- **Contexto:** Las representaciones latentes de muchos aspectos y la base del comportamiento humano están naturalmente incrustadas en la expresión de las emociones que se encuentran en el habla humana. Esto significa la importancia de extraer datos de audio recopilados de conversaciones humanas para extraer emociones humanas. La capacidad de capturar y representar las emociones humanas será una característica importante de la inteligencia artificial de próxima generación, con la expectativa de una interacción más estrecha con los humanos.

- **Metodología:** La metodología propuesta se basa en el enfoque de extracción de características basado en Bag of Audio Words (BoAW) y un modelo de red neuronal recurrente de última generación. el modelo RNN adoptado captura la información contextual a largo y corto plazo en la conversación. En la Figura 8 se observa como el modelo funciona en tres fases. (1) Extracción de características, (2) Incorporación de características de Bolsa de palabras de audio (BoAW) y (3) Extracción de emociones El diagrama general se presenta en la Figura 2.2.

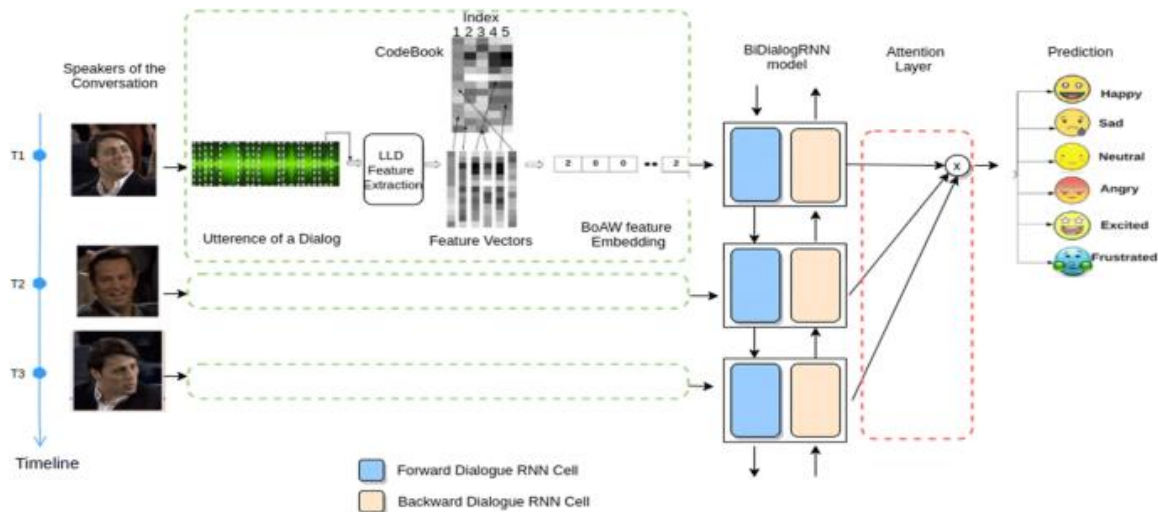


Figura 2.2 Modelo de red neuronal recurrente basado en bolsa de palabras de audio para la detección de emociones a partir de audio [25]

- **Resultados:** La evaluación de la nueva característica propuesta realizada sobre el conjunto de datos IEMOCAP arroja resultados prometedores de 60,87% de precisión ponderada (aproximadamente 20% de mejora) y 60,97% de precisión no ponderada en el reconocimiento de las 6 emociones básicas que superan a los modelos actuales de última generación que utilizan la modalidad de audio.

Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters (Calificaciones de confianza y reconocimiento de emociones predichas por el tipo de estímulo vocal y los parámetros prosódicos) [27]

- **Contexto:** El lenguaje humano no solo comunica significado a través de la semántica, sino que también transmite emociones mediante atributos vocales como el tono y el volumen. En las investigaciones sobre el reconocimiento de emociones en la voz, se observa una considerable variabilidad en los tipos de estímulos utilizados y en los procedimientos empleados para analizar su impacto en las emociones. Se ha argumentado, además, que una metacognición precisa favorece interpretaciones correctas y seguras en las tareas de reconocimiento emocional. Sin embargo, estas conexiones apenas han sido exploradas anteriormente. Se busca abordar esta laguna investigando cómo el tipo de estímulo vocal y los atributos prosódicos del habla influyen en el reconocimiento de emociones y en la confianza de una persona en una respuesta específica.

- **Metodología:** Se analizó el material de estímulo con respecto a parámetros asociados a la frecuencia (frecuencia fundamental media (F0), F0 mínima, F0 máxima, desviación estándar F0, fluctuación), parámetros relacionados con la energía/amplitud (brillo, amplitud [dB], pico amplitud, HNR media, HNR máxima, desviación estándar HNR), y características temporales (duración, tiempo pico),

empleando para ello paquetes de script GSU.

• **Resultados:** Los que se presentan en la Tabla 2.11 indican que el reconocimiento de emociones a partir de explosiones vocales depende mucho más de la preservación de la información acústica (amplitud, tono). Aunque los parámetros relacionados con la amplitud, la frecuencia y el tiempo son importantes en la inflexión emocional del lenguaje hablado, la proporción de varianza explicada por estas medidas es ligeramente menor. para oraciones o estímulos con una connotación semántica negativa o positiva.

Tabla 2.11 Matriz de confusión, comparaciones de error de clasificación por cosque aleatorio (RF) y juicio de los oyentes sobre categorías de emociones y IC del 95% de la prueba binomial exacta para errores de clasificación de los oyentes [27].

Tipo de estimulación	Emociones	Juicio emocional							Errores de clasificación	
		Enojo	Miedo	Felicidad	Neutral	Tristeza	Asco	Sorpres a	Oyentes (IC 95%)	RF
Oraciones Neutrales	Enojo	97.28	0.10	1.17	0.79	0.00	0.17	0.49	2.72%(2.16 % 44.86%)	60 %
	Miedo	3.86	56.97	8.62	5.35	0.86	1.03	23.21	43,03% [41,22%; 44,86%]	35 %
	Felicidad	2.76	1.00	76.00	2.59	0.83	0.55	16.27	24,00% [22,46%; 25,60%]	75 %
	Neutral	4.86	0.76	0.38	86.17	6.59	0.62	0.62	13,83% [12,60%; 15,14%]	15 %
	Tristeza	0.28	12.69	0.03	9.31	77.31	0.17	0.21	22,69% [21,18%; 24,26%]	30 %
	Asco	2.10	4.86	0.48	10.55	13.28	67.52	1.21	32,48% [30,78%; 34,22%]	65 %
	Sorpresa	-	-	-	-	-	-	-	-	-

Perceptual Cues in Nonverbal Vocal Expressions of Emotion (Señales perceptivas en expresiones vocales no verbales de emoción) [28]

- **Contexto:** El trabajo sobre las expresiones faciales de las emociones y el habla con inflexiones emocionales ha delineado con éxito algunas de las propiedades físicas que subyacen al reconocimiento de las emociones. Para identificar las señales acústicas utilizadas en la percepción de expresiones emocionales no verbales como la risa y los gritos.

- **Metodología:** Primero, los estímulos emocionales se categorizaron y calificaron para establecer que los oyentes pudieran identificar y calificar los sonidos de manera confiable y proporcionar matrices de confusión. En segundo lugar, se midieron las propiedades acústicas de amplitud, tono y perfil espectral de los estímulos. Un procedimiento de análisis discriminante estableció que estas medidas acústicas proporcionaban suficiente discriminación entre expresiones de categorías emocionales para permitir una clasificación estadística precisa.

- **Resultados:** Al igual que las señales afectivas en las expresiones faciales y el habla con inflexiones emocionales, el carácter emocional percibido de las vocalizaciones afectivas se puede predecir en función de sus características físicas.

Detection and Analysis of Emotion From Speech Signals (Detección y análisis de emociones a partir de señales del habla) [26]

- **Contexto:** Este artículo realiza una investigación experimental sobre el reconocimiento de emociones a través del habla humana. Inicialmente, se examinó la discernibilidad de las características emocionales en el habla, seguido por la clasificación de las emociones en un conjunto de datos personalizado. Este proceso de clasificación se llevó a cabo utilizando diversos clasificadores.

- **Metodología:** Se utilizaron datos de tres sujetos diferentes en un estudio que abordó el reconocimiento de emociones a través del habla. Durante la recopilación de datos, se grabaron las voces de tres mujeres distintas que expresaron emociones específicas mientras se registraba su discurso. Estos sujetos, de origen ruso, pronunciaron palabras en ruso en diferentes estados emocionales. La grabación del discurso se llevó a cabo con un teléfono móvil posicionado a unos 15 cm de la boca de los participantes, en un dormitorio convencional de $25m^2$.

Para extraer las características de los segmentos de voz grabados, se emplearon funciones de MATLAB, con especial énfasis en los coeficientes cepstrales de frecuencia Mel (MFCC), reconocidos por su eficacia en experimentos de clasificación de audio. Los MFCC capturan la información espectral a corto plazo de la señal de habla, proporcionando datos relevantes sobre la calidad de la voz y los efectos de producción. Este proceso implica la transformada del coseno del

logaritmo real del espectro de energía de corto plazo, seguido de la representación en escala de frecuencia Mel. Además, después del preénfasis, los segmentos del discurso se someten a la apertura en ventanas utilizando la ventana de Hamming, diseñada para reducir el efecto de fuga y prevenir discontinuidades en el dominio del tiempo durante la Transformada Rápida de Fourier.

En términos de clasificación emocional, se amplió el conjunto de datos a 30 sujetos, con edades comprendidas entre los 20 y 45 años, abarcando una proporción igual de hombres y mujeres. Se grabó la voz de cada sujeto en 30 ocasiones, y los datos resultantes se etiquetaron en tres categorías de emociones: neutral, ira y alegría. Este enfoque se adoptó en lugar de incluir todas las diferentes emociones humanas, buscando así una clasificación más específica y significativa.

• **Resultados:** Primero se estudió el efecto del estado emocional al pronunciar palabras simples. Se tomaron las distancias máximas para estudiar el efecto. Los resultados trazados para un solo sujeto se muestran en la Figura 2.3.

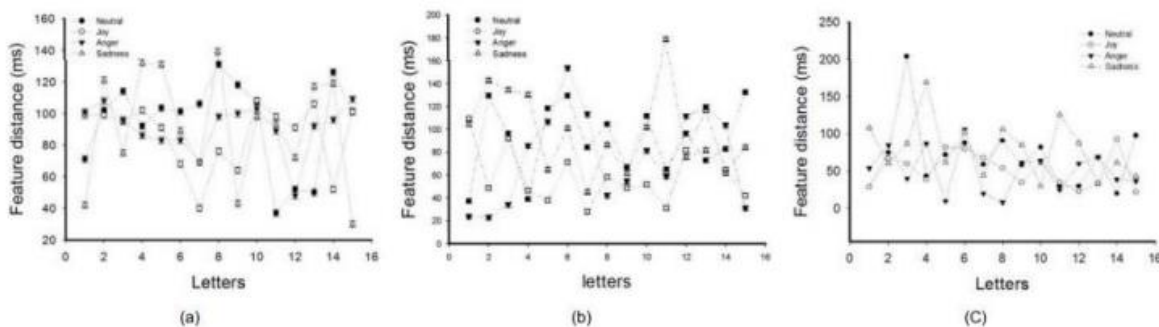


Figura 2.3 Las distancias entre los picos de letras individuales para (a) sujeto 1, (b) sujeto 2 y (c) sujeto 3 [26].

Al observar los resultados del grafico se observa que la emoción tristeza tiene una mayor distancia en comparación con otros, seguida por la emoción neutral, siendo el enojo el segundo con menor distancia y la alegría con la distancia más pequeña. El estudio de las distancias entre características se realizó también para las palabras. Los gráficos obtenidos se trazaron en la Figura 2.4. Aquí también se observó un comportamiento similar.

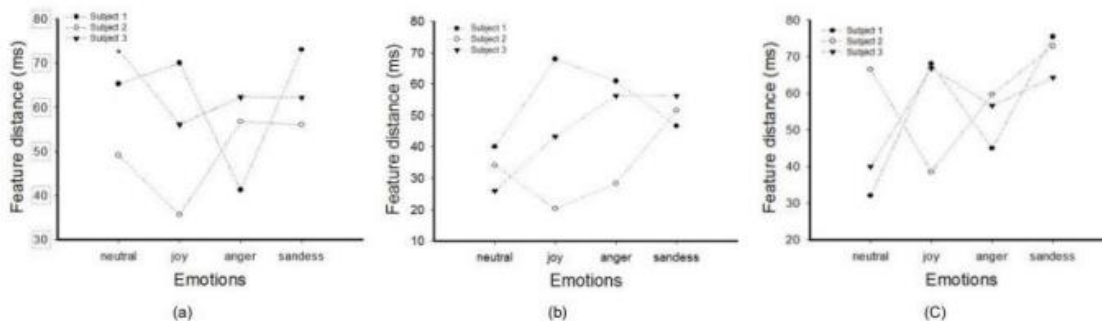


Figura 2.4 Las distancias entre los picos de la señal del habla de (a) palabra 1, (b) palabra 2 y (c) palabra 3 para todos los sujetos bajo cuatro emociones: neutral, alegría, ira y tristeza [26].

Después de comparar la pronunciación de las palabras y las letras separadas contenidas en esas palabras en las Figuras 2.3 y 2.4, se puede inferir que el estado emocional de una persona tiene una influencia directa sobre las señales vocales que produce independientemente de lo que se dice.

2.3 Análisis de artículos

La revisión realizada muestra que, aunque no existe una base extensa de estudios enfocados directamente en la detección de clonación de voz, las investigaciones sobre reconocimiento vocal, clasificación de género, identificación de edad y análisis emocional aportan elementos valiosos para comprender el potencial de las características acústicas en tareas de identificación.

Los trabajos demuestran que la voz humana es una fuente rica de información, en la que rasgos como la frecuencia fundamental, la prosodia, el timbre y otros descriptores pueden revelar datos sobre la identidad o el estado del hablante, incluso en entornos con variabilidad de ruido y grabación.

El análisis de artículos centrados en emociones, género o edad evidencia que las técnicas de inteligencia artificial, especialmente el aprendizaje profundo, han permitido avances significativos en la extracción de patrones complejos de la señal de voz. Sin embargo, también revela que los estudios existentes se orientan principalmente a clasificar atributos biológicos o emocionales, dejando poco explorada la identificación de manipulaciones o falsificaciones, como ocurre en el caso de las voces clonadas mediante inteligencia artificial. Esta carencia de literatura específica subraya la necesidad de investigaciones que profundicen en descriptores acústicos capaces de detectar alteraciones sutiles producidas durante el proceso de síntesis o clonación.

Por ello, aunque los artículos analizados no aborden la clonación de voz de manera directa, su revisión resulta esencial para sentar bases conceptuales y metodológicas. Comprender cómo se han empleado técnicas de análisis de señal, extracción de características y modelos de clasificación en áreas afines permite identificar estrategias que pueden adaptarse o extenderse al problema de la detección de voces sintéticas. De este modo, este capítulo no solo ofrece una panorámica del reconocimiento vocal en campos relacionados, sino que también justifica la pertinencia y originalidad de esta tesis, la cual busca cubrir un vacío evidente en la literatura científica al proponer un análisis específico

para diferenciar voces naturales de voces clonadas.

CAPÍTULO 3 MARCO TEÓRICO

El capítulo explica la biología de la voz, describiendo el papel de las cuerdas vocales, el fuelle (pulmones y músculos respiratorios) y los resonadores (faringe, boca y cavidad nasal) en la producción del sonido. Luego aborda la modelación digital de la voz, mostrando cómo la señal analógica se convierte en digital mediante muestreo y cuantificación. Posteriormente, se presentan las métricas de evaluación (matriz de confusión, exactitud, recall, precisión y F1) para medir el desempeño de modelos de clasificación. Finalmente, se detallan las librerías y entornos de programación utilizados, como Python con Numpy, Pandas, Librosa y Sklearn, entre otros.

3.1 Biología de la voz

Normalmente cuando se habla de la voz, se piensa en el cuello, pero, realmente la voz tanto hablada como cantada se produce debido a la coordinación de las estructuras del aparato respiratorio, digestivo y de un conjunto de músculos, formando las ondas que se conocen como voz [44].

A continuación, se describen las tres principales estructuras encargadas de la producción de la voz [44].

3.1.1 Las cuerdas vocales (laringe)

Las cuerdas vocales se localizan en la laringe y, aunque su función principal es la protección de las vías respiratorias, han sido adaptadas para la producción de sonido. Estas estructuras forman parte del vibrador del aparato fonador y generan la voz mediante su vibración cuando el aire proveniente de los pulmones pasa a través de ellas. La frecuencia y regularidad de esta vibración determinan la frecuencia fundamental de la voz, característica acústica estrechamente relacionada con el tono vocal y ampliamente utilizada en el análisis y caracterización de señales de voz.

3.1.2 Fuelle

La segunda parte, denominada fuelle, está constituida por los pulmones, la caja torácica, el músculo diafragma y los músculos del abdomen. Este conjunto de estructuras es el encargado de proporcionar la presión del aire necesaria para la vibración de las cuerdas vocales, permitiendo regular la intensidad y la duración del sonido emitido.

3.1.3 Resonadores

La tercera parte nombrada los resonadores, se conforma por la faringe, la boca y la cavidad nasal, su forma y posición determina el timbre de nuestra voz, se puede variar el tamaño de la boca moviendo la posición de la lengua (Figura 3.1- C), de la mandíbula (Figura 3.1-D), de los labios (Figura 3.1-B) o del paladar (Figura 3.1-G).

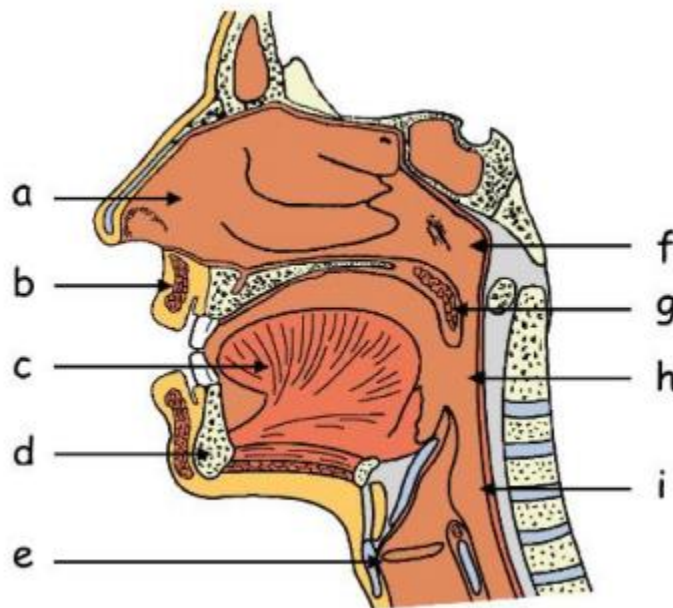


Figura 3.1 Corte sagital de la cabeza. a: cavidad nasal, b: labio, c: lengua, d: mandíbula, e: laringe, f: porción nasal de la faringe, g: velo del paladar, h: porción oral de la faringe, i: porción laríngea de la faringe [44].

3.2 Modelación de la señal digital de voz

La vibración de un objeto (la cual se denomina fuente del sonido) causa que las moléculas de aire alrededor de dicho elemento empiecen a moverse y propagarse. Esta propagación en un medio es lo que ocasiona el sonido. Para el caso de la voz, las cuerdas vocales vibran por el aire expulsado de los pulmones. Posteriormente, los elementos de la cavidad vocal modifican las vibraciones para que cuando salgan de nuestros labios, los oídos humanos perciban la voz.

Esencialmente la voz humana es una señal analógica que lleva consigo información de estas características. Este fenómeno se modela mediante un sistema de tiempo discreto [45], [46], [47].

El proceso de conversión de una señal analógica (véase la Figura 3.2-a) a digital consta de 2 etapas, muestreo y cuantificación. El primer proceso es el muestreo y consiste en una señal muestreada a intervalos de tiempo uniformes (véase las Figuras 3.2-b y 3.2-c). Por ejemplo, un muestreo de 2000 hercios significa que se tomaron 2000 muestras de amplitud por segundo. El segundo proceso consiste en tomar el valor de cada amplitud y cuantificarlo en un número limitado de bits. Cuanto mayor sea el número de bits, mejor calidad de audio se obtendrá (véase la Figura 3.2-d).

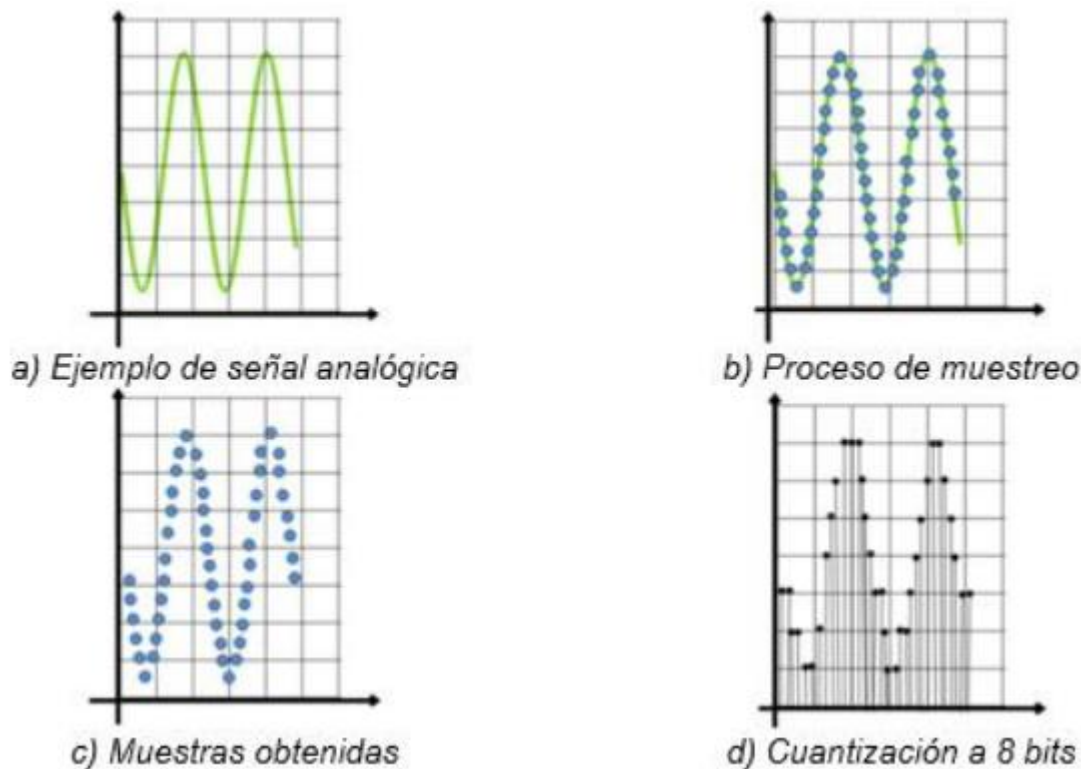


Figura 3.2 a) Ejemplo de señal analógica, b) Proceso de muestreo, c) Muestras obtenidas, d) Cuantización a 8 bits [15].

3.3 Frecuencia fundamental (F0)

La frecuencia fundamental (f_0) es un parámetro acústico que representa la frecuencia con la que vibran las cuerdas vocales durante la producción del sonido. Este componente es uno de los indicadores más importantes en el análisis de la voz, ya que permite describir tanto características fisiológicas del hablante como aspectos expresivos y lingüísticos de su habla. Asimismo, la F_0 está estrechamente relacionada con la percepción del tono de la voz, por lo que desempeña un papel fundamental en la comunicación verbal.

Desde el punto de vista acústico, la frecuencia fundamental corresponde a la fuente sonora generada por la vibración de las cuerdas vocales. Dicha señal se propaga a través del tracto vocal, cuyas configuraciones actúan como un sistema de filtrado que da forma al sonido final emitido por el hablante. Durante este proceso, la f_0 no permanece constante, sino que presenta variaciones continuas que dependen tanto del contenido lingüístico como del estado emocional de la persona que habla. Estas variaciones contribuyen de manera importante a la entonación y a la expresividad del habla en la comunicación cotidiana [48].

3.4 Métricas de evaluación

Para el cálculo de las métricas se utiliza la matriz de confusión que como se observa en [49] son utilizadas para evaluar el rendimiento de un modelo predictivo. Ver tabla 3.1. Se definen las categorías: S: = Sintética y H: = Humana.

Se realizan las predicciones y se contabilizan respecto a la pertenencia real para cada instancia.

Tabla 3.1 Matriz de confusión del problema de clasificación [49].

	Realidad humana	Realidad sintética
Predicción humana	TP	FP
Predicción sintética	FN	TN

Donde:

- TP + FP + TN + FN: = Total de predicciones realizadas.
- TP + TN: = Total de predicciones correctas.
- FP + FN: = Total de predicciones incorrectas.
- TP + FP: = Total de predicciones de humanas.
- FN + TN: = Total de predicciones de sintéticas.
- TP + FN: = Total de humanas.
- FP + TN: = Total de sintéticas.

Las fórmulas de las métricas son las siguientes:

$$\text{Exactitud (Accuracy)} = \frac{TP + TN}{TP + TN + FP + FN} \quad 1$$

$$\text{Precisión (Precision)} = \frac{TP}{TP + FP} \quad 2$$

$$\text{Exhaustividad (Recall)} = \frac{TP}{TP + FN} \quad 3$$

$$\text{Especificidad (Specificity)} = \frac{TN}{TN + FP} \quad 4$$

$$\text{Sensibilidad (Sensitivity)} = \frac{TP}{TP + FN} \quad 5$$

$$\text{Puntaje F1 (F1-score)} = 2 * \frac{\text{Precisión} * \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}} \quad 6$$

- La exactitud es la proporción de predicciones correctas sobre todas las predicciones.
- La precisión es la proporción de predicciones correctas de voces humanas de todo el conjunto de predicciones de voces humanas.

- La exhaustividad es la proporción de predicciones correctas de voces humanas entre el conjunto de voces sintéticas.
- La especificidad es la proporción de predicciones correctas de voces sintéticas entre todo el conjunto de predicciones de voces sintéticas.
- La sensibilidad es idéntica a la exhaustividad.
- La medida F1 es la media armónica entre la exhaustividad y la precisión. Es usada cuando la cantidad de elementos en las categorías no están equilibradas.

3.5 Librerías y entornos de desarrollo

En el desarrollo del sistema se utilizaron diferentes librerías las cuales se listan en la Tabla 3.2.

Tabla 3.2 Especificaciones de hardware, versión de Python y librerías utilizadas.

Librería	Versión	Descripción
Matplotlib	3.9.2	Librería de visualización para crear gráficos en 2D.
Numpy	1.23.4	Librería para cálculos numéricos.
Pandas	1.5.3	Librería para análisis y manipulación de datos.
Librosa	0.10.2.post1	Librería para el procesamiento de audios.
CUDA	11.2.67	Librería para hacer uso de GPU.
Crepe	0.0.16	Librería para calcular F0 de los audios.
sklearn	1.6.1	Librería para el uso de máquina vector soporte.

3.6 Algoritmo CREPE

Un algoritmo para el seguimiento del tono que se fundamenta en el análisis de datos mediante una red neuronal convolucional profunda que trabaja directamente con la forma de onda en el dominio del tiempo. Se plantea como una alternativa superior a metodologías previas como pYIN, que combinan procesamiento de señales digitales (DSP) y heurísticas, destacando que CREPE alcanza resultados de vanguardia, superando o igualando a pYIN en rendimiento y mostrando una mayor resistencia ante el ruido debido a esto se consideró como el algoritmo a utilizar para la extracción de características.

Se vale de una red neuronal convolucional profunda que analiza señales de audio en el dominio del tiempo para la estimación del tono como se muestra en la Figura 3.3. La entrada es un extracto de 1024 muestras en el dominio del tiempo. señal de audio, utilizando una frecuencia de muestreo de 16 kHz. Hay seis capas convolucionales que dan como resultado una representación de 2048 dimensiones, que se conecta densamente a la capa de salida con activaciones sigmoideas correspondientes a un vector de salida de 360 dimensiones.

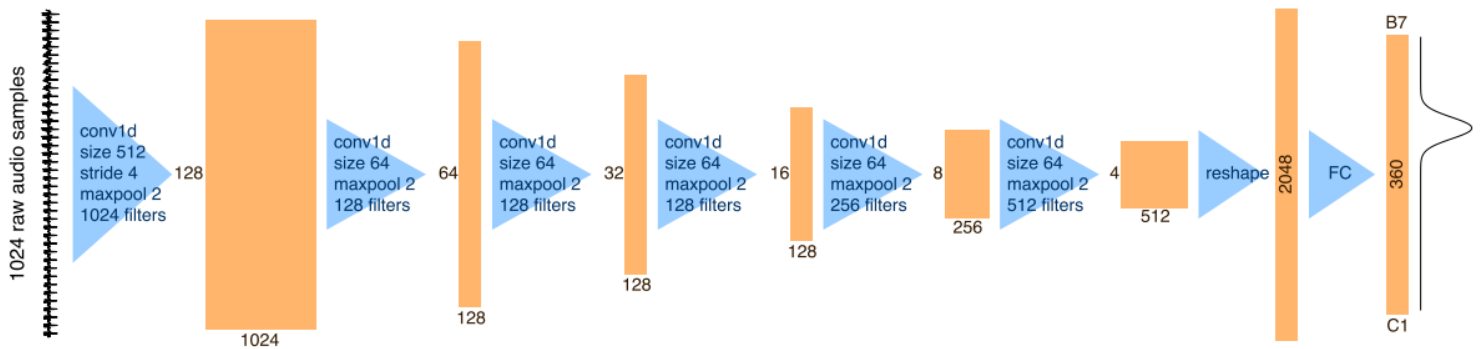


Figura 3.3 La arquitectura del rastreador de tono CREPE. Las seis capas convolucionales operan directamente sobre la señal de audio en el dominio del tiempo, produciendo un vector de salida que se aproxima a una curva gaussiana, que luego se utiliza para derivar la estimación exacta del tono [50].

3.7 Filtros

Los filtros son herramientas fundamentales en el procesamiento de señales de audio, ya que permiten modificar el contenido espectral de una señal mediante la amplificación, atenuación o preservación de determinadas frecuencias dentro del espectro audible. Su función principal es mejorar la calidad de la señal, reducir interferencias no deseadas y resaltar la información relevante para su posterior análisis o procesamiento.

En el ámbito del audio, los filtros se utilizan para limpiar regiones específicas del espectro de frecuencias, ya sea eliminando componentes indeseables como el ruido, o reforzando aquellas bandas que contienen información importante de la señal. Estas operaciones son indispensables en aplicaciones como el análisis acústico, los sistemas de grabación, la producción musical, el procesamiento digital de señales y los sistemas de comunicación.

3.7.1 Tipos de filtros

Los filtros pueden clasificarse según la forma en que permiten o restringen el paso de determinadas bandas de frecuencia [51] [52].

3.7.1.1 Filtro paso bajo (Low Pass, LP)

Los filtros seleccionan frecuencias bajas hasta la frecuencia de corte f_c y atenúan frecuencias superiores a f_c , como se muestra en la figura 3.4.

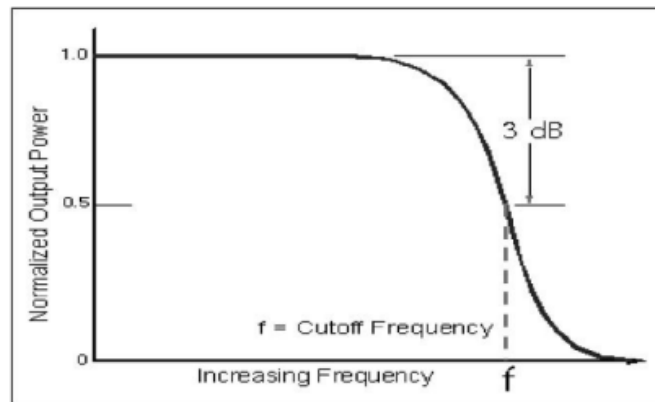


Figura 3.4 Filtro pasa baja [52]

3.7.1.2 Filtro paso alto (High Pass, HP)

Los filtros seleccionan frecuencias superiores a f_c y atenúan las frecuencias inferiores a f_c , como se muestra en la figura 3.5.

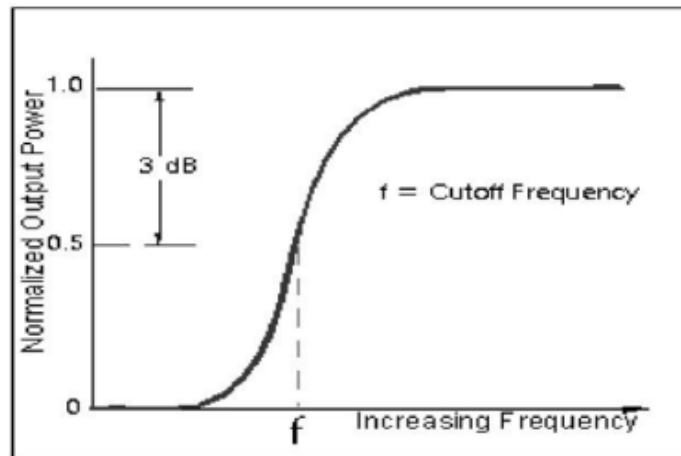


Figura 3.5 Filtro pasa alto [52]

3.7.1.3 Filtro pasa banda (Band Pass, BP)

Los filtros seleccionan frecuencias entre una frecuencia de corte inferior (f_{cl}) y una frecuencia de corte superior (f_{ch}). Las frecuencias inferiores a f_{cl} y superiores a f_{ch} se atenúan, como se muestra en la figura 3.6.

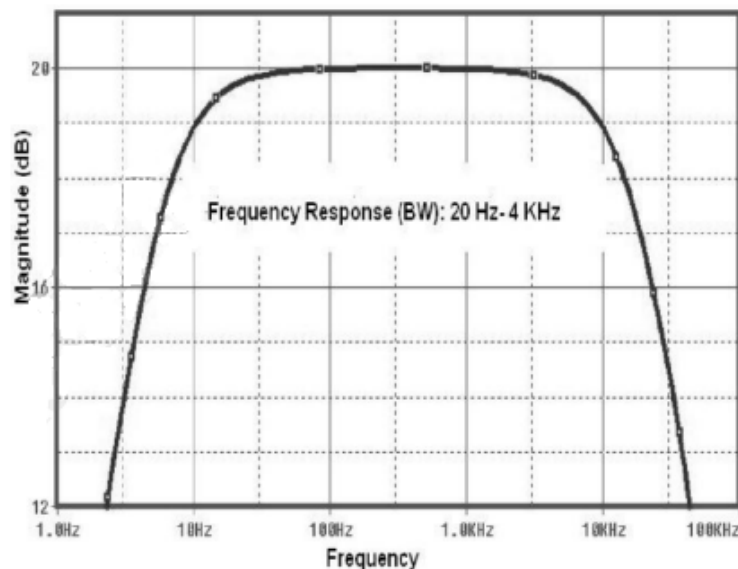


Figura 3.6 Filtro pasa banda [52]

3.8 XgBoost [53]

Es un método de aprendizaje automático utilizado en distintos problemas de clasificación, regresión y *ranking*. Este enfoque se basa en la construcción secuencial de árboles de decisión, donde cada nuevo árbol se ajusta considerando los errores cometidos por los anteriores. Dentro de este enfoque se encuentra el *gradient tree boosting*, técnica utilizada en diversas aplicaciones prácticas.

En este contexto, XGBoost es un sistema de *tree boosting* diseñado como una solución escalable de principio a fin. Este sistema incorpora un algoritmo sensible a la dispersión de los datos (*sparsity-aware*), así como un método de *weighted quantile sketch* para el aprendizaje aproximado de los árboles de decisión. Además, su implementación considera aspectos relacionados con los patrones de acceso a caché, la compresión de datos y la fragmentación (*sharding*), con el objetivo de facilitar el manejo de grandes volúmenes de información.

3.9 Máquina vector soporte (SVM)

Las Máquinas de Vectores de Soporte (SVM, del inglés *Support Vector Machine*) son modelos utilizados para la resolución de problemas de clasificación, regresión, agrupamiento y multiclase. Dentro de la tarea de clasificación, las SVM pertenecen a la categoría de los clasificadores lineales, ya que inducen separadores lineales, también llamados hiperplanos [54].

Estos hiperplanos pueden construirse directamente en el espacio original de los datos cuando los ejemplos son linealmente separables o cuasi separables, o bien en un espacio transformado de características cuando no lo son. La búsqueda del hiperplano de separación en estos espacios transformados, normalmente de alta dimensión, se realiza de forma implícita mediante el uso de las funciones kernel.

En este contexto, cada entrada de dos dimensiones puede transformarse en un espacio de características de diez dimensiones con el objetivo de encontrar un hiperplano capaz de separar los ejemplos transformados. La frontera de decisión lineal asociada a dicho hiperplano se transforma en un límite de decisión polinomial de grado tres en el espacio de entradas. Un pequeño aumento en la dimensionalidad del espacio de entrada puede provocar un gran aumento en la dimensionalidad del espacio de características. En el caso límite, incluso pueden existir espacios de características de dimensión infinita.

Por esta razón, el problema de optimización se expresa únicamente en su forma dual, ya que la solución no depende de la dimensionalidad del espacio de características, sino de la cardinalidad del conjunto de vectores soporte. Sin embargo, cuando la transformación del espacio de entradas al espacio de características se define a partir de un conjunto infinito de funciones base, surge la necesidad de transformar ejemplos de dimensión finita hacia espacios de dimensión infinita, situación que se aborda a través del uso de funciones kernel.

Por definición, una función kernel es una función $K : X \times X \rightarrow \mathbb{R}$ que asigna a cada par de elementos del espacio de entrada X , un valor real correspondiente al producto escalar de las imágenes de dichos elementos en un nuevo espacio f (espacio de características) [54].

Kernel	Fórmula	
Lineal	$K_L(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$	7
Polinómico de grado-p	$K_P(\mathbf{x}, \mathbf{x}') = [\gamma \langle \mathbf{x}, \mathbf{x}' \rangle + \tau]^p, \gamma > 0$	8
Gaussiano	$K_G(\mathbf{x}, \mathbf{x}') = \exp \left(-\gamma \ \mathbf{x} - \mathbf{x}'\ ^2 \right) \equiv \exp \left(-\gamma \langle \mathbf{x} - \mathbf{x}', \mathbf{x} - \mathbf{x}' \rangle \right), \gamma > 0$	9
Sigmoidal	$K_S(\mathbf{x}, \mathbf{x}') = \tanh(\gamma \langle \mathbf{x}, \mathbf{x}' \rangle + \tau)$	10

CAPÍTULO 4 ANÁLISIS Y DISEÑO DEL SISTEMA

En este capítulo se explica el diseño del sistema, junto con los elementos presentes en los respectivos ambientes. También, se detallan las técnicas utilizadas en los módulos: preprocesamiento de audios y conjunto de datos creados; fases empleadas para efectuar las experimentaciones que se exponen en el capítulo 5.

4.1 Metodología

La metodología general se compone de tres etapas: procesamiento, extracción de características y clasificación. A partir de esta estructura común, se desarrollaron cuatro variantes experimentales, las cuales difieren principalmente en la forma de extraer, representar y analizar las características acústicas de los audios.

Como se muestra en [50], [55], se observa que el algoritmo CREPE es ampliamente utilizado y recomendado para la extracción tonal de la voz. Este algoritmo ha sido empleado en aplicaciones como la clonación de voces [9], [10], debido a su alta precisión en la estimación de la frecuencia fundamental. Por ello, en el presente trabajo se propone hacer uso de dicho algoritmo para la extracción de las características utilizadas en la clasificación de voces clonadas y voces naturales.

A continuación, se muestra en la Figura 4.1 el esquema general de la metodología propuesta para la solución de la problemática.

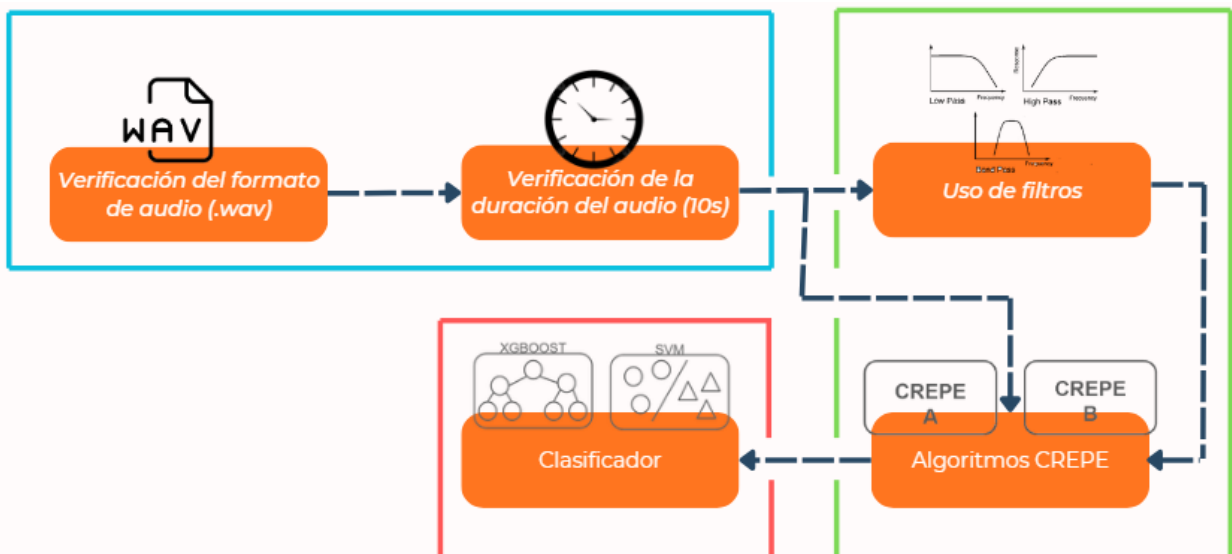


Figura 4.1 Esquema general de metodología.

4.2 Técnicas

4.2.1 Procesamiento

- Todos los audios fueron normalizados y segmentados a una duración estándar de 10 segundos, garantizando uniformidad para el análisis posterior.
- Se verifica que todos los archivos estén en formato .WAV, transformando aquellos que no cumplieran esta condición mediante scripts en Python.
- Para una las variantes, se incorporó la aplicación de filtros digitales de tipo *Butterworth* de quinto orden para resaltar componentes relevantes de la señal vocal y reducir el ruido.

4.2.2 Extracción de características:

- Se emplearon los siguientes filtros: filtro pasa baja con atenuación de frecuencias superiores a 500 Hz, filtro pasa alta con atenuación de frecuencias inferiores a 2000 Hz y, filtro pasa banda que permite el paso de frecuencias entre 500 Hz y 2000 Hz, dichos filtros serán aplicados después de la normalización de tiempo y antes del procesamiento para extraer las características.
- En [13] se señala que existe una diferencia significativa entre la frecuencia fundamental (F_0) promedio de hombres y mujeres. En términos generales, la F_0 promedio es notablemente más alta en mujeres y el rango de F_0 en Hz es más amplio en comparación con el de los hombres. A pesar de que las diferencias en el rango de F_0 entre sexos desaparecen cuando se expresan en semitonos o como un factor de modulación, los valores promedio de F_0 continúan siendo distintos. En resumen, la F_0 promedio de los hombres no es equivalente al F_0 en las mujeres.
- El factor de modulación, en el contexto de la voz, se refiere a la variación de la frecuencia fundamental (F_0) de la voz a lo largo de un período de tiempo. Esta medida indica cuánto fluctúa la frecuencia fundamental alrededor de su valor promedio. La modulación de frecuencia puede ser cuantificada en porcentaje o en unidades de frecuencia (como Hz). Un alto factor de modulación sugiere que la voz presenta numerosas variaciones en tono, lo cual puede interpretarse como expresividad o variabilidad vocal.

Debido a las diversas formas de extraer las características se desarrollaron cuatro variantes de la metodología. En todas las variantes se utilizó el algoritmo CREPE para la extracción de la frecuencia fundamental (F_0), aplicando distintas estrategias de análisis y considerando diferentes versiones del modelo según la implementación utilizada:

- En la primera variante, se empleó el algoritmo CREPE (versión *Tiny*) implementado en MATLAB [14] donde se calcula hF_0 (estimación de frecuencia fundamental híbrida) y se utiliza el vector de valores para generar un histograma.
- En la segunda variante, también utilizando CREPE en MATLAB, se calculó la desviación estándar de la F_0 por segmento, considerando valores con y sin ceros, como una representación estadística de la variabilidad tonal. En ambas variantes basadas en MATLAB, se utilizó la versión *Tiny* del modelo CREPE, dado que es la

única disponible en dicha implementación.

- En la tercera variante, se usó CREPE implementado en Python, generando un vector de características directamente a partir de los valores de F0 extraídos.
- En la cuarta variante, se utilizaron las versiones de CREPE implementadas en Python. Se extrajo un vector de características desde la F0, el cual fue empleado en una estrategia de clasificación distinta, utilizando el algoritmo XGBoost. En Python se pueden encontrar distintas variantes de CREPE y se evaluaron las versiones *Tiny*, *Small*, *Medium*, *High* y *Full*, lo que permitió analizar el impacto del tamaño del modelo sobre la precisión de la estimación de la frecuencia fundamental.

4.2.3 Clasificación:

En las variantes uno, dos y tres, los vectores de características fueron clasificados mediante una Máquina de Vectores de Soporte (SVM). En la cuarta variante, se utilizó el algoritmo XGBoost para la clasificación debido a la gran cantidad de datos y los tiempos de entrenamiento. En todos estos casos, se aplicó una división del conjunto de datos en un 80% para entrenamiento y 20% para validación.

4.2.4 Filtros

Con el objetivo de mejorar la calidad de los datos de entrada y facilitar una extracción de características más precisa, se emplearon filtros en la etapa de preprocesamiento de las señales de voz. Esta decisión se fundamentó en observaciones previas, donde se identificó que las frecuencias ubicadas en rangos bajos de Hertz aportaban más información para distinguir entre voces clonadas y naturales. En este sentido, y con base en criterio propio, se definieron los siguientes rangos de filtrado: un filtro pasa baja con atenuación de frecuencias superiores a 500 Hz, un filtro pasa alta que atenúa las frecuencias inferiores a 2000 Hz, y un filtro pasa banda que permite el paso de frecuencias entre 500 Hz y 2000 Hz. La aplicación de estos filtros permitió aislar regiones espectrales específicas de interés, eliminando componentes menos relevantes o potencialmente ruidosos. Dichos filtros se aplicaron después de la normalización temporal y antes del proceso de extracción de características, con el fin de optimizar la representación de la señal para el análisis posterior.

4.2.5 Conjuntos de datos

La sección se integra del detalle de los conjuntos de voces, tanto naturales como artificiales que se utilizan en el proyecto de tesis.

4.2.6 Conjuntos de datos de voces naturales

Common Voice [56]

Common Voice es un proyecto desarrollado por Mozilla con el objetivo de construir un conjunto de datos de voz abierto, diverso e inclusivo, orientado a mejorar la accesibilidad y la representación en las tecnologías de reconocimiento de voz. Este recurso se conforma a partir de las contribuciones voluntarias de personas que graban y validan frases en distintos idiomas.

Para este trabajo, se seleccionó un subconjunto de audios de *Common Voice* equivalente en número a los audios clonados generados con cada herramienta, con el fin de mantener un equilibrio entre las clases. Asimismo, se procuró que la selección incluyera una proporción equitativa de voces masculinas y femeninas. La mayoría de los audios presenta algún grado de ruido de fondo, aunque también se incluyeron algunos provenientes de ambientes controlados, aunque en menor cantidad.

Principales Características:

1. **Datos Abiertos** Las grabaciones recopiladas se distribuyen bajo la licencia CC0 (dominio público), lo que garantiza su disponibilidad sin restricciones. Esto permite que los datos sean utilizados en investigación, desarrollo de *software* y en la implementación de tecnologías de voz.
2. **Idiomas y Diversidad** El proyecto incluye soporte para más de 100 idiomas y está diseñado para capturar una amplia gama de acentos y dialectos, lo que contribuye a reflejar la diversidad lingüística y cultural a nivel global.
3. **Contribuciones Voluntarias** La plataforma permite que cualquier persona participe leyendo frases para grabar su voz o revisando las grabaciones de otros para validar su calidad. Este enfoque colaborativo es fundamental para la construcción del conjunto de datos.
4. **Propósito del Proyecto** La iniciativa busca democratizar el acceso a las tecnologías de voz, reduciendo sesgos en los sistemas existentes y promoviendo herramientas inclusivas que representen a una mayor variedad de usuarios.

4.2.7 Conjunto de datos de voces generados artificialmente

Tal como se menciona en la sección 3, el conjunto de datos ASVspoof 2019 [57] ha sido empleado como base para los experimentos relacionados con la detección de voces sintéticas. Sin embargo, este conjunto presenta ciertas limitaciones ya descritas, entre las que destacan la presencia de un ritmo constante, entonación monótona y una calidad vocal artificial o robotizada. Estas características hacen que las voces clonadas contenidas en

dicho conjunto sean fácilmente distinguibles. Por esta razón, se planteó la necesidad de generar nuevos audios utilizando las herramientas descritas en el capítulo 1, con el propósito de disponer de datos más realistas y exigentes para los sistemas de detección evaluados.

Por consiguiente, se llevó a cabo la generación de un nuevo conjunto de datos compuesto por aproximadamente 1300 audios clonados, tal como se presenta en la Tabla 4.1. Para este fin, se emplearon cuatro herramientas distintas de generación de voz humana, seleccionadas debido a las variadas arquitecturas que caracterizan a dichas plataformas. Asimismo, se consideró la inclusión de aplicaciones tanto de pago como gratuitas. Todos los audios fueron generados utilizando frases diferentes entre sí, evitando la repetición de contenido idéntico.

Tabla 4.1 Conjunto de voces generadas

Conjunto	Tiempo de audios	Número de audios
Voces Clonadas Español	149 minutos	796
Voces Clonadas Inglés	97 minutos	460

Los audios que integran el *dataset* son de tiempos variables, desde los 11 segundos hasta los 2 minutos, en las Tablas 4.2, 4.3 ,4.4 y 4.5 detallan la información de los conjuntos de datos generados, por herramienta.

Tabla 4.2 Conjunto de voces clonadas con la Herramienta RVC

Idioma	Género	Cantidad
Español	Mujer	100
	Hombre	100
	Subtotal	200
Inglés	Mujer	60
	Hombre	100
	Subtotal	160
Total		360

Tabla 4.3 Conjunto de voces clonadas Herramienta: Applio

Idioma	Género	Cantidad
Español	Mujer	56
	Hombre	100
	Subtotal	156
Inglés	Mujer	64
	Hombre	140
	Subtotal	204
Total		360

Tabla 4.4 Conjunto de voces clonadas Herramienta: Elevenlabs

Idioma	Género	Cantidad
Español	Mujer	60
	Hombre	40
	Subtotal	100
Inglés	Mujer	40
	Hombre	60
	Subtotal	100
Total		200

Tabla 4.5 Conjunto de voces clonadas Herramienta: Voice.ia

Idioma	Género	Cantidad
Español	Mujer	40
	Hombre	60
	Subtotal	100
Inglés	Mujer	200
	Hombre	36
	Subtotal	236
Total		336

Librerías

En el desarrollo del sistema se utilizaron diferentes librerías las cuales se listan en la Tabla 4.6.

Tabla 4.6 Especificaciones de hardware, versión de Python y librerías utilizadas

Librería	Versión	Descripción
Matplotlib	3.9.2	Librería de visualización para crear gráficos en 2D.
Numpy	1.23.4	Librería para cálculos numéricos.
Pandas	1.5.3	Librería para análisis y manipulación de datos.
Librosa	0.10.2.post1	Librería para el procesamiento de audios.
CUDA	11.2.67	Librería para hacer uso de GPU.
Crepe	0.0.16	Librería para calcular F0 de los audios.
sklearn	1.6.1	Librería para el uso de máquina vector soporte.

CAPÍTULO 5 EXPERIMENTACIÓN Y RESULTADOS

En este capítulo se detallan las configuraciones previas a la experimentación (hiperparámetros y distribución de las particiones). Se evalúan las cinco variantes de la metodología, por ello se tienen cinco casos de prueba; se presenta el análisis de los resultados obtenidos recurriendo a las métricas de evaluación y en base a ello, se opta por el mejor modelo predictivo y se exporta para su uso en los respectivos ambientes de ejecución.

5.1 Experimentación

5.1.1 Procesado de los audios

Se verifica que los archivos de audio estén en formato “.WAV”, por lo que el primer paso consiste en transformar aquellos que no cumplen con este requisito.

A continuación, se aplican los filtros de “pasa baja”, “pasa alta” y “pasa banda” [58] a los audios en formato “.WAV”.

Posteriormente para estimar la frecuencia fundamental (pitch) de una señal monofónica, se utiliza un enfoque basado en la autocorrelación de la señal y el análisis en sub-bandas de frecuencia. El procedimiento para obtener un histograma que representa la distribución de las frecuencias fundamentales estimadas a lo largo del tiempo es el siguiente:

1. Segmentación del audio en frames temporales:
La señal de audio se divide en ventanas temporales superpuestas (*frames*) de duración fija (por ejemplo, 50 ms con salto de 10 ms), lo que permite analizar la evolución temporal del *pitch*.
2. Cálculo de la autocorrelación:
Para cada *frame*, se calcula la función de autocorrelación de la señal completa y también para cada una de las sub-bandas filtradas en distintas bandas de frecuencia (por ejemplo, entre 50 Hz y 800 Hz, segmentadas en 8 sub-bandas). La autocorrelación resalta las periodicidades presentes en la señal, que son indicativas de la frecuencia fundamental.
3. Extracción de picos en la autocorrelación.
La función de autocorrelación para cada *frame* exhibe picos en retardos que corresponden a posibles periodos fundamentales. Se extraen estos picos y se transforman en frecuencias invirtiendo el retardo ($\text{frecuencia} = \text{frecuencia de muestreo} / \text{retardo}$).

4. Estimación del *pitch* por *frame* usando redes neuronales convolucionales.
Para mejorar la robustez y precisión, se utiliza un modelo de red neuronal convolucional entrenado para identificar, entre las sub-bandas, la mejor estimación del *pitch* para cada *frame*. Este modelo procesa la autocorrelación de las sub-bandas y determina la clase (sub-banda) con mayor probabilidad de contener la frecuencia fundamental.
5. Construcción del vector de *pitch*.
Se genera un vector con la frecuencia fundamental estimada para cada *frame*. Este vector contiene valores de *pitch* en Hz para cada instante temporal analizado.
6. Construcción del histograma del *pitch*.
Finalmente, para obtener un histograma, se considera el conjunto de valores del vector de *pitch* (excluyendo valores cero o no válidos). El histograma representa la distribución de las frecuencias fundamentales a lo largo del audio, mostrando la cantidad de veces que cada rango de frecuencia ocurre. Esto facilita visualizar qué frecuencias son dominantes en la señal analizada.

5.2 Variante uno

5.2.1 Procesado de los audios

Teniendo en cuenta lo anterior, se usó el modelo de *Crepe-Tiny* de Matlab [14] para el procesamiento de hF0. Este algoritmo puede ser clasificado como una red poco profunda debido a su estructura simplificada y al reducido número de capas que la conforman. A diferencia del modelo original CREPE, que incorpora seis capas convolucionales seguidas de una capa de salida, este algoritmo consta únicamente de dos capas convolucionales y una capa softmax final. Esta diferencia en la profundidad implica una menor cantidad de transformaciones jerárquicas de la señal, lo cual limita la complejidad de patrones que puede aprender, pero a su vez mejora la eficiencia computacional y reduce el riesgo de sobreajuste, especialmente en escenarios con recursos limitados o conjuntos de datos reducidos.

Este algoritmo es un rastreador de tono monofónico basado en una red neuronal convolucional poco profunda que opera sobre la función de autocorrelación normalizada en el dominio del tiempo. El código original ha sido modificado para permitir el procesamiento de múltiples archivos de audio, ya que su versión base estaba diseñada para trabajar únicamente con un archivo a la vez. Además, se adaptó para almacenar los histogramas generados y los vectores de datos extraídos, funcionalidades que no estaban contempladas en la implementación original.

Los datos de los histogramas se dividen en los canales “R”, “G” y “B”, registrando la frecuencia de los valores en un rango de 0 a 255 dentro de un archivo de Excel. Finalmente, estos valores son utilizados en una máquina de soporte vectorial, creando un conjunto de datos para cada filtro por separado.

En la Figura 5.1 se puede observar un ejemplo de cómo el histograma muestra en el eje X el tiempo en segundos, desde los 0 hasta los 10, en color se muestra la potencia y, en el eje Y se muestra la frecuencia de la voz.

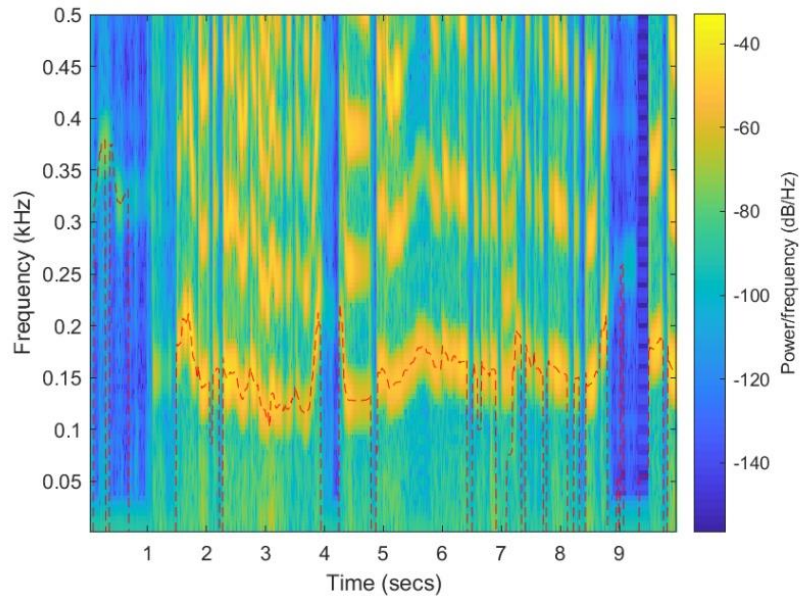


Figura 5.1 Ejemplo del histograma obtenido de un audio de voz natural.

5.2.2 Clasificación

Se seleccionó el uso de Máquinas de Vectores de Soporte (SVM) debido a su alta capacidad para resolver problemas de clasificación binaria en espacios de alta dimensión [54], como el caso de la representación mediante histogramas de F0 [14]. Se evaluaron los kernels “lineal, polinomial, gaussiano y sigmoide”, haciendo uso de los hiper parámetros por defecto en la librería sklearn. Del conjunto de datos se utilizó el 80% para entrenamiento y el 20% para validación. En la Tabla 5.1, se observan los resultados con la aplicación de los filtros en combinación con los kernels utilizados.

Tabla 5.1 Resultados de la primera variante con la métrica exactitud.

Características con filtros	MVS-Kernel	Exactitud
Pasa banda	Lineal	0.9018
Pasa banda	Polinomial	0.8661
Pasa banda	RBF	0.9018
Pasa banda	Sigmoideo	0.8304
Pasa alto	Lineal	0.8571
Pasa alto	Polinomial	0.8750
Pasa alto	RBF	0.8571
Pasa alto	Sigmoideo	0.8393
Pasa baja	Lineal	0.9018
Pasa baja	Polinomial	0.8482
Pasa baja	RBF	0.8571
Pasa baja	Sigmoideo	0.9554
Sin filtro	Lineal	0.9018
Sin filtro	Polinomial	0.8839
Sin filtro	RBF	0.9018
Sin filtro	Sigmoideo	0.9196

El análisis de los resultados obtenidos mediante la combinación de diferentes filtros y kernels de SVM revela patrones significativos que pueden guiar la selección del método más adecuado para tareas específicas de clasificación de señales.

- Filtro Pasa Banda: El filtro pasa banda muestra un rendimiento consistente con los kernels lineal y RBF, ambos alcanzando una exactitud de 0.9018. Sin embargo, el kernel RBF logra una sensibilidad perfecta (1.0000), lo que sugiere una capacidad superior para identificar correctamente las señales positivas, aunque con una especificidad moderada. Este kernel podría ser preferido en aplicaciones donde es crucial minimizar los falsos negativos.
- Filtro Pasa Alta: En el caso del filtro pasa alta, el kernel polinomial presenta un buen equilibrio entre exactitud (0.8750) y alta sensibilidad (0.9286), pero con una especificidad constante de 0.8214. Esto indica que el kernel polinomial es eficiente en la detección de verdaderos positivos, aunque su capacidad para discriminar los verdaderos negativos es limitada. Este filtro y kernel pueden ser útiles en escenarios donde la detección correcta de señales es prioritaria sobre la especificidad.
- Filtro Pasa Baja: El filtro pasa baja combinado con el kernel sigmoide destaca al lograr la mayor exactitud (0.9554) y alta especificidad (0.9286), lo que sugiere una excelente capacidad tanto para identificar correctamente las señales positivas como para minimizar los falsos positivos. Este método sería ideal para aplicaciones que requieren una alta exactitud general y una excelente discriminación entre señales positivas y negativas.
- Sin Filtro: Sin la aplicación de un filtro, el kernel sigmoide también muestra un rendimiento notable, con una exactitud de 0.9196 y una alta especificidad (0.9464). Esto sugiere que este método es altamente efectivo para aplicaciones donde se necesita un análisis detallado de las características de frecuencia de las señales.

5.2.3 Conclusión

Los resultados obtenidos en este estudio demuestran que las características extraídas del histograma de la frecuencia fundamental (F0) son una herramienta eficaz para diferenciar entre voces clonadas y naturales. El uso de un clasificador basado en Máquinas de Vectores de Soporte (SVM) permitió alcanzar una precisión del 95%, lo que indica un alto nivel de acierto en la identificación de voces generadas artificialmente. Estos hallazgos validan la utilidad de la frecuencia fundamental como una característica distintiva en el análisis de señales de voz y refuerzan su potencial en aplicaciones de autenticación y detección de fraudes en entornos digitales. Con estos resultados, se concluye que la descripción ha cumplido satisfactoriamente, destacando la relevancia de esta metodología en el estudio y la seguridad de los sistemas de clonación de voz.

5.3 Segunda variante

5.3.1 Procesado de los audios

Se verifica que los archivos de audio estén en formato “.WAV”; por ello, el primer paso consiste en convertir al formato adecuado aquellos que no lo cumplan. Para esto, se utiliza un código en Python que recorre todos los archivos de audio en la carpeta y, si es necesario, realiza la conversión correspondiente.

Luego, los audios se normalizan a una duración de 10 segundos para asegurar que todos tengan la misma longitud.

5.3.2 Extracción de características

Una vez normalizados, se emplea el código de CREPE en MATLAB [59] para obtener el vector de frecuencias fundamentales a lo largo del tiempo. A partir de este vector, se calculan dos valores: la desviación estándar considerando todo el muestreo y otra desviación estándar que ignora los valores cero, los cuales suelen estar relacionados con pausas o errores en la detección.

Estos datos se utilizan como entrada para un clasificador SVM, que se prueba con distintos tipos de núcleos: lineal, polinomial, RBF (función base radial) y sigmoideo como se muestra en la Tabla 5.2.

Tabla 5.2 Resultados de la segunda variante con la métrica exactitud.

Kernel	Exactitud
Lineal	0.5989
RBF	0.6286
Polinomial	0.5685
Sigmoide	0.479

5.3.3 Conclusión

Según los resultados obtenidos, el clasificador SVM con el núcleo RBF presentó la mayor precisión, con un valor de 0.6286. El núcleo lineal obtuvo una precisión de 0.5989, seguido por el núcleo polinomial con 0.5685. El desempeño más bajo se registró con el núcleo sigmoide, que alcanzó una precisión de 0.479.

Estos valores indican que el uso directo de la desviación estándar como única característica en el vector de datos no proporciona resultados satisfactorios en términos de precisión del modelo.

5.4 Tercera variante

5.4.1 Procesado de audios

Se diseñó un enfoque técnico orientado a la extracción y análisis de características acústicas de señales de audio, empleando para ello la herramienta CREPE (*Convolutional Representation for Pitch Estimation*) [50], reconocida por su capacidad para estimar con precisión el tono fundamental mediante redes neuronales convolucionales.

Esta metodología tuvo como objetivo principal evaluar la eficacia del análisis de estas

características en la clasificación de muestras de voz, diferenciando entre voces reales y voces clonadas generadas mediante técnicas de síntesis artificial. Como paso inicial, se procedió a la normalización de todos los archivos de audio utilizados, ajustando cada uno a una duración estándar de 10 segundos. Esta etapa fue esencial para garantizar la uniformidad temporal de las señales, reduciendo la variabilidad no controlada y permitiendo condiciones homogéneas para el análisis posterior.

5.4.2 Extracción de características

Se emplearon los vectores de características generados por el algoritmo CREPE, los cuales comprenden principalmente dos variables: "*frequency*", que representa la estimación del tono fundamental de la voz en hercios (Hz) a lo largo del tiempo, y "*confidence*", que refleja la probabilidad de presencia del tono en cada instante de la señal. Ambas variables fueron utilizadas como atributos para el entrenamiento de un modelo de clasificación basado en máquinas de vectores de soporte (SVM, por sus siglas en inglés).

Las métricas fueron obtenidas a partir de ambos vectores generados por CREPE: la frecuencia estimada y el nivel de confianza asociado a cada muestra. Ambos valores fueron utilizados de forma conjunta como entrada para la Máquina de Vectores de Soporte (SVM), permitiendo que el modelo aprendiera tanto de las variaciones tonales como del grado de certeza que CREPE asignaba a cada estimación. Esta combinación buscó mejorar la capacidad del sistema para identificar patrones asociados a voces clonadas, considerando no solo el contenido tonal sino también la estabilidad y credibilidad de la señal.

Con el objetivo de evaluar el desempeño del modelo bajo diferentes configuraciones, se realizaron entrenamientos utilizando diversos núcleos o kernels: lineal, radial básico (RBF), polinomial y sigmoide. En todos los casos, se utilizaron los parámetros por defecto ofrecidos por la implementación estándar del clasificador. Los resultados obtenidos para cada una de estas variantes se presentan en la Tabla 5.3, donde se comparan los indicadores de rendimiento alcanzados por el modelo en la tarea de clasificación entre voces reales y voces clonadas.

Tabla 5.3 Resultados de la tercera variante con la métrica exactitud.

Kernel	Exactitud
Lineal	0.5469
RBF	0.4526
Polinomial	0.6685
Sigmoide	0.573

5.4.3 Conclusión

Los resultados obtenidos al emplear una Máquina de Vectores de Soporte (SVM) con parámetros por defecto, utilizando como entrada las columnas de frecuencia y nivel de confianza generadas por CREPE, muestran variaciones relevantes en el desempeño según el núcleo utilizado. El mejor resultado se alcanzó con el núcleo polinomial, con una exactitud de 0.6685, seguido por el sigmoide (0.573) y el lineal (0.5469). El núcleo RBF presentó el menor rendimiento con una exactitud de 0.4526. Estos resultados sugieren que, bajo esta configuración, la relación no lineal capturada por el núcleo polinomial permite una mejor separación entre las clases evaluadas. No obstante, los valores obtenidos indican que esta representación de los datos, aunque útil, podría beneficiarse de una optimización adicional

en el preprocesamiento o en la selección de características.

5.5 Cuarta variante

5.5.1 Procesado de audios

Para esta variante de la metodología se diseñó un enfoque orientado a la extracción y análisis de características de señales de audio utilizando la herramienta CREPE (*Convolutional Representation for Pitch Estimation*) [50], complementado con un proceso de segmentación estadística, con el objetivo de evaluar su eficacia en la clasificación de voces reales y voces clonadas.

El primer paso consistió en normalizar todos los audios a una duración estándar de 10 segundos. Una vez normalizados, cada audio fue procesado mediante tres tipos de filtros: pasa baja, pasa banda y pasa alta, permitiendo analizar el comportamiento de las frecuencias bajas, medias y altas por separado. Esta etapa es fundamental para identificar qué rangos de frecuencia contienen más información relevante para la tarea de clasificación.

5.5.2 Extracción de características

Se empleó CREPE para extraer vectores de características relacionadas con el *pitch*, utilizando un muestreo temporal de 0.01 segundos, lo que generó 1000 valores por archivo de audio. Este nivel de detalle permitió capturar variaciones sutiles en el tono que podrían ser indicativas de manipulación digital o síntesis artificial.

La salida de CREPE consiste en un vector que contiene dos componentes por cada muestra: la frecuencia fundamental estimada (en Hz) y un valor de confianza, que indica el grado de certeza del modelo respecto a la estimación realizada en ese instante. Estos dos valores permiten no solo capturar la evolución tonal de la señal, sino también identificar posibles regiones inestables o de baja calidad.

Con los vectores obtenidos, se calcularon diferentes configuraciones de desviación estándar (DE), como medida de la variabilidad tonal:

- Una única DE global calculada sobre la secuencia completa de frecuencias (1000 valores).
- Desviaciones estándar cada 5 muestras, generando 200 segmentos de 0.05 segundos.
- Desviaciones estándar cada 10 muestras, con 100 segmentos de 0.1 segundos.

Las métricas fueron obtenidas a partir de ambos vectores generados por CREPE: la frecuencia estimada y el nivel de confianza asociado a cada muestra. Se utilizaron distintas configuraciones del modelo, incluyendo las variantes disponibles en CREPE según el tamaño del modelo (*full*, *medium*, *small*, entre otros) y la opción de aplicar o no suavizado sobre las predicciones. Estos valores se emplearon de forma conjunta como entrada para el algoritmo XGBoost, que mostró un mejor desempeño respecto a otras alternativas.

Para el entrenamiento se utilizaron los siguientes parámetros: *n_estimators* = 200 (número de árboles), *max_depth* = 6 (profundidad máxima por árbol), *learning_rate* = 0.1 (tasa de aprendizaje), *tree_method* = 'hist_gpu' (construcción de árboles mediante GPU), y *random_state* = 42 (semilla para reproducibilidad).

La configuración permitió que el modelo aprovechara tanto las variaciones tonales como el

nivel de confianza asignado por CREPE para identificar patrones vinculados a voces clonadas, los resultados se muestran en las Tablas 5.4, 5.5, 5.6, 5.7 y 5.8.

Tabla 5.4 Resultados de la cuarta variante, con el algoritmo *full*, con la métrica de precisión.

Algoritmo	Suavizado	Filtro	CSV	Precisión
<i>Full</i>	Sin suavizado	Pasa alto	valores cada 10.csv	0.68157981
			valores cada 3.csv	0.66008981
			valores cada 5.csv	0.66572945
		Pasa baja	valores cada 10.csv	0.76117877
			valores cada 3.csv	0.74823224
			valores cada 5.csv	0.74903634
		Pasa banda	valores cada 10.csv	0.74483481
			valores cada 3.csv	0.71104419
			valores cada 5.csv	0.71929666
		Todo	valores cada 10.csv	0.81530343
			valores cada 3.csv	0.79900714
			valores cada 5.csv	0.80444985
	Suavizado	Pasa alto	valores cada 10.csv	0.68413332
			valores cada 3.csv	0.6623161
			valores cada 5.csv	0.66940632
		Pasa baja	valores cada 10.csv	0.78258721
			valores cada 3.csv	0.75123721
			valores cada 5.csv	0.75823653
		Pasa banda	valores cada 10.csv	0.72725576
			valores cada 3.csv	0.69937277
			valores cada 5.csv	0.70711687
		Todo	valores cada 10.csv	0.76376504
			valores cada 3.csv	0.7364144
			valores cada 5.csv	0.742085

Tabla 5.5 Resultados de la cuarta variante, con el algoritmo *large*, con la métrica de precisión.

Algoritmo	Suavizado	Filtro	CSV	Precisión
<i>Large</i>	Sin suavizado	Pasa alto	valores cada 10.csv	0.72683936
			valores cada 3.csv	0.69612441
			valores cada 5.csv	0.70273132
		Pasa baja	valores cada 10.csv	0.7622645
			valores cada 3.csv	0.75578494
			valores cada 5.csv	0.75466841
		Pasa banda	valores cada 10.csv	0.75515873
			valores cada 3.csv	0.74096838
			valores cada 5.csv	0.74237753
		Todo	valores cada 10.csv	0.74745961
			valores cada 3.csv	0.73877685
			valores cada 5.csv	0.74436783
	Suavizado	Pasa alto	valores cada 10.csv	0.7395791
			valores cada 3.csv	0.7198091
			valores cada 5.csv	0.72302957
		Pasa baja	valores cada 10.csv	0.78503294
			valores cada 3.csv	0.76618768
			valores cada 5.csv	0.76760016
		Pasa banda	valores cada 10.csv	0.75772643
			valores cada 3.csv	0.74685503
			valores cada 5.csv	0.74788429
		Todo	valores cada 10.csv	0.7390819
			valores cada 3.csv	0.72131513
			valores cada 5.csv	0.72389882

Tabla 5.6 Resultados de la cuarta variante, con el algoritmo *medium*, con la métrica de precisión.

Algoritmo	Suavizado	Filtro	CSV	Precisión
<i>Medium</i>	Sin suavizado	Pasa alto	valores cada 10.csv	0.69528875
			valores cada 3.csv	0.67449154
			valores cada 5.csv	0.67440287
		Pasa baja	valores cada 10.csv	0.75448718
			valores cada 3.csv	0.74363022
			valores cada 5.csv	0.74382702
		Pasa banda	valores cada 10.csv	0.7403808
			valores cada 3.csv	0.72016127
			valores cada 5.csv	0.72298774
		Todo	valores cada 10.csv	0.74930297
			valores cada 3.csv	0.72837419
			valores cada 5.csv	0.74180568
	Suavizado	Pasa alto	valores cada 10.csv	0.71148393
			valores cada 3.csv	0.69693307
			valores cada 5.csv	0.69878924
		Pasa baja	valores cada 10.csv	0.77827833
			valores cada 3.csv	0.75358183
			valores cada 5.csv	0.75584453
		Pasa banda	valores cada 10.csv	0.74460577
			valores cada 3.csv	0.7245611
			valores cada 5.csv	0.72355885
		Todo	valores cada 10.csv	0.75202632
			valores cada 3.csv	0.72518076
			valores cada 5.csv	0.73141392

Tabla 5.7 Resultados de la cuarta variante, con el algoritmo *small*, con la métrica de precisión.

Algoritmo	Suavizado	Filtro	CSV	Precisión
<i>Small</i>	Sin suavizado	Pasa alto	valores cada 10.csv	0.68012629
			valores cada 3.csv	0.65514344
			valores cada 5.csv	0.65709212
		Pasa baja	valores cada 10.csv	0.74858836
			valores cada 3.csv	0.74135205
			valores cada 5.csv	0.73898213
		Pasa banda	valores cada 10.csv	0.73084205
			valores cada 3.csv	0.70547584
			valores cada 5.csv	0.71412756
		Todo	valores cada 10.csv	0.74392802
			valores cada 3.csv	0.7273956
			valores cada 5.csv	0.73115907
	Suavizado	Pasa alto	valores cada 10.csv	0.68168712
			valores cada 3.csv	0.65409332
			valores cada 5.csv	0.65936533
		Pasa baja	valores cada 10.csv	0.77555071
			valores cada 3.csv	0.74922822
			valores cada 5.csv	0.75246051
		Pasa banda	valores cada 10.csv	0.73073427
			valores cada 3.csv	0.70845808
			valores cada 5.csv	0.71388336
		Todo	valores cada 10.csv	0.74650878
			valores cada 3.csv	0.71694372
			valores cada 5.csv	0.72523269

Tabla 5.8 Resultados de la cuarta variante, con el algoritmo *tiny*, con la métrica de precisión.

Algoritmo	Suavizado	Filtro	CSV	Precisión
<i>Tiny</i>	Sin suavizado	Pasa alto	valores cada 10.csv	0.66966026
			valores cada 3.csv	0.65230617
			valores cada 5.csv	0.65325847
		Pasa baja	valores cada 10.csv	0.74059699
			valores cada 3.csv	0.72220274
			valores cada 5.csv	0.7225763
		Pasa banda	valores cada 10.csv	0.71676185
			valores cada 3.csv	0.70060168
			valores cada 5.csv	0.70219466
		Todo	valores cada 10.csv	0.73123949
			valores cada 3.csv	0.72525062
			valores cada 5.csv	0.72752462
	Suavizado	Pasa alto	valores cada 10.csv	0.67031452
			valores cada 3.csv	0.65703362
			valores cada 5.csv	0.65834771
		Pasa baja	valores cada 10.csv	0.75564563
			valores cada 3.csv	0.72719755
			valores cada 5.csv	0.73076923
		Pasa banda	valores cada 10.csv	0.72185635
			valores cada 3.csv	0.69969094
			valores cada 5.csv	0.70302552
		Todo	valores cada 10.csv	0.73112033
			valores cada 3.csv	0.70161646
			valores cada 5.csv	0.70589023

Conclusión

En función de los resultados obtenidos, se observó que la mejor precisión se alcanzó utilizando el modelo *full* de CREPE sin suavizado, sin filtrado adicional y con valores tomados cada 10 muestras, logrando una precisión de 0.8153. Le siguieron en desempeño las configuraciones *small* con suavizado y paso bajo (0.7756), y *large* con las mismas condiciones (0.7850), ambas también con muestreo cada 10 valores. Las variantes *medium* y *tiny*, aun con suavizado y filtrado pasa baja, mostraron una precisión inferior (0.7558 y 0.7556 respectivamente), incluso con un muestreo más denso en el caso de *medium* (cada 5 valores). Estos resultados sugieren que, en este contexto específico, el modelo *full* sin suavizado ni filtros adicionales ofrece una representación más eficaz de las características relevantes para la tarea de clasificación.

Resultados finales

Con base en los resultados obtenidos en las distintas variantes evaluadas, se observa que la frecuencia fundamental (f_0) constituye una característica relevante para la diferenciación entre voces naturales y voces clonadas; sin embargo, su efectividad depende directamente de la forma en que es representada y procesada. La Variante 1 mostró el mejor desempeño, evidenciando que la representación de f_0 mediante histogramas proporciona una descripción más informativa para la clasificación que el uso directo de valores aislados o estadísticas simples. En contraste, las Variantes 2 y 3 presentaron precisiones moderadas, lo que indica que el uso directo del vector de datos o de las salidas básicas de CREPE limita la capacidad de separación entre clases sin un tratamiento previo más robusto. Por su parte, la Variante 4 evidenció que la selección del modelo de extracción de f_0 y la configuración del muestreo influyen de manera significativa en el desempeño del clasificador, destacando que el modelo *full* de CREPE sin filtrado adicional ofreció los mejores resultados dentro de esta configuración. En conjunto, los resultados permiten concluir que la combinación adecuada entre el método de extracción de características, su representación y el algoritmo de clasificación es determinante para lograr una detección más precisa de voces clonadas.

CAPÍTULO 6 CONCLUSIONES

En esta sección, se exponen las conclusiones generales del proyecto de tesis, la justificación de los objetivos y alcances logrados. También se anexan las aportaciones y actividades académicas adicionales hechas durante el programa de posgrado.

6.1 Conclusiones generales

Los resultados obtenidos indican que las métricas de la media y el valor mínimo presentan una diferencia significativa entre los conjuntos de voces naturales y clonadas. Este enfoque también sugiere la viabilidad de trabajar con múltiples voces, aunque con la limitación de que solo una persona puede hablar a la vez para garantizar la precisión del análisis.

En conjunto, los distintos enfoques evaluados en este estudio evidencian que las características acústicas extraídas a partir de la frecuencia fundamental (f_0), ya sea mediante análisis estadístico, representación en histogramas o vectores generados por CREPE A, permiten distinguir de forma efectiva entre voces naturales y clonadas. Entre las estrategias analizadas, el uso del modelo *full* de CREPE B sin suavizado ni filtrado adicional, combinado con un clasificador XGBoost, alcanzó la mayor precisión (0.8153), superando a otras configuraciones y modelos. Asimismo, el análisis mediante histogramas y SVM logró una precisión del 95%, destacando la utilidad de esta representación para tareas de clasificación.

Por otro lado, se identificó que el uso exclusivo de estadísticas simples como la desviación estándar o la media, si bien permite evidenciar diferencias entre clases, no es suficiente por sí solo para alcanzar niveles elevados de precisión. Del mismo modo, se observaron variaciones importantes en el desempeño de los distintos núcleos del clasificador SVM, dependiendo de la configuración de los vectores de entrada. En particular, se obtuvo un mejor resultado con el núcleo polinomial (0.6685) al utilizar directamente las salidas de CREPE A.

En términos generales, los hallazgos respaldan la eficacia del uso de características tonales, especialmente cuando se aprovechan modelos de mayor complejidad o representaciones más estructuradas de la señal, para abordar tareas de detección de voz clonada. Sin embargo, también se reconocen limitaciones, como la necesidad de analizar una sola voz a la vez para garantizar la fiabilidad del modelo. Este trabajo contribuye así a la comprensión del comportamiento de distintas metodologías frente al mismo problema y establece una base para futuras mejoras en la identificación automática de voces generadas artificialmente.

En la Tabla 6.1 se muestran los mejores resultados de las diversas variantes evaluadas.

Tabla 6.1 Mejores resultados de todas las variantes.

Variante	Modelo	Mejor resultado
Primera variante	SVM	95%
Segunda variante	SVM	62%
Tercera variante	SVM	66%
Cuarta variante	XGBoost	81%

6.1.1 Cumplimiento de los compromisos

En las Tablas 6.2, 6.3 y 6.4 se listas los entregables comprometidos, el cumplimiento de los alcances y objetivos.

Tabla 6.2 Objetivos de investigación

Objetivo	Estado	Observaciones
Desarrollo de un modelo para identificar voz natural vs. sintética (es/en)	Cumplido	Modelos funcionales desarrollados utilizando dos algoritmos CREPE para la extracción de la frecuencia fundamental como característica principal de la voz. Asimismo, se emplearon dos clasificadores para evaluar el rendimiento del sistema, obteniendo resultados cuantitativos en cada variante. El mejor desempeño alcanzó un 95% de precisión mediante el uso de histogramas.
Objetivos específicos		
Estudio del estado del arte sobre caracterización de voz	Cumplido	Revisión del estado del arte documentada en los cuatro reportes entregados a lo largo de la maestría, identificando que la frecuencia fundamental constituye una característica relevante para diferenciar voces clonadas de voces naturales.
Análisis de herramientas de IA para clonación de voz	Cumplido	Se realizó un análisis de técnicas y herramientas actuales utilizadas en la clonación de voz. Como parte de este objetivo, se elaboró y publicó el artículo "Review of AI-generated Voice Cloning Tools", en el que se describen las cinco herramientas más relevantes del área [60].
Recolección de conjuntos de audios naturales y sintéticos	Cumplido	Reunidos desde plataformas públicas y herramientas de IA.
Análisis de técnicas de extracción de características	Cumplido	Se evaluaron dos enfoques (CREPE, histogramas).
Implementación de algoritmo de clasificación	Cumplido	Se probaron SVM y XGBoost con diversas configuraciones.
Evaluación cuantitativa del rendimiento	Cumplido	Se reportaron métricas como exactitud en varios escenarios.

Tabla 6.3 Alcances planteados

Alcance	Estado	Observaciones
Audios en ambiente no controlado	Cumplido	El sistema fue probado con grabaciones con ruido ambiental moderado.
Audios en reconocimiento de voz natural y sintética (español e inglés) ambiente no controlado	Cumplido	Se incluyeron ambos idiomas en los conjuntos de prueba.
Evaluación con al menos 2 herramientas de clonación gratuitas	Cumplido	Se usaron tanto herramientas de paga como gratuitas.
Modelo preciso y eficiente en recursos	Cumplido	Se logró buena precisión; optimización haciendo uso de las versiones “Tiny”.

Tabla 6.4 Entregables de investigación

Entregable	Estado	Observaciones
Reporte del estado del arte	Cumplido	Incluye revisión de técnicas de clonación y detección de voz.
Reportes semestrales y de resultados	Cumplido	Se entregó el presente informe con resultados y análisis.
Análisis de herramientas de clonación de voz	Cumplido	Se entregó un artículo sobre las herramientas en CITCA el día 21 de mayo
Código del sistema	Parcialmente cumplido	Se cuenta con el código sin interfaz
Conjuntos de audios naturales y sintéticos	Parcialmente cumplido	Se cuenta con el conjunto de datos generados, pero no se publica
Publicación de artículo	Cumplido	Se entregó un artículo en COMIA el día 15 de mayo con el título “Sistema de reconocimiento de voz humana y sintética” con ISSN:1870-4069 [61] y en CITCA el día 21 de mayo con el nombre “Review of AI-generated Voice Cloning Tools” con ISSN:2992-8338 [60].
Tesis	Cumplido	El presente documento es la versión final de la tesis.

6.2 Contribuciones del trabajo

Este trabajo aporta diversas contribuciones al estudio de la detección de voces clonadas, tanto a nivel metodológico como experimental. En primer lugar, se exploraron y adaptaron distintas metodologías para la identificación de voces sintetizadas, utilizando como base el algoritmo CREPE. Tras una revisión del estado del arte, no se identificó un uso previo de

esta herramienta como extractor de características para la clasificación entre voces reales y clonadas, lo cual representa una aportación novedosa dentro de este campo.

Asimismo, se implementó un esquema de preprocesamiento acústico mediante filtros pasa baja, pasa alta y pasa banda, cuyos rangos fueron definidos a criterio propio. Estos filtros permitieron resaltar distintos componentes espectrales de la señal, facilitando el análisis del impacto de cada rango de frecuencia en la diferenciación entre voces naturales y sintetizadas. Dicho preprocesamiento se aplicó antes del cálculo de la frecuencia fundamental (f_0), la cual fue utilizada como vector representativo de cada muestra de audio.

También se construyó un conjunto de datos balanceado, compuesto por voces clonadas generadas a través de distintas herramientas contemplando diversidad de arquitecturas, plataformas gratuitas y de paga, así mismo las voces naturales extraídas del proyecto *Common Voice*. Se garantizó una distribución equitativa entre voces masculinas y femeninas, además de incluir grabaciones tanto con ruido de fondo como en condiciones controladas.

Finalmente, se aplicaron distintos núcleos del clasificador SVM para evaluar la capacidad discriminativa de los vectores de f_0 , lo que permitió analizar el comportamiento de esta representación frente a la tarea de clasificación binaria entre voces reales y clonadas.

TRABAJOS A FUTURO

Como trabajos a futuro, se propone integrar el modelo dentro de una aplicación funcional y desarrollar una versión optimizada para dispositivos móviles. Asimismo, se plantea explorar el uso de otros algoritmos de extracción de frecuencia fundamental, como “*Harvest*”, y analizar nuevas características relevantes de la voz que puedan mejorar la identificación de falsificaciones vocales.

6.3 Actividades académicas

- Primer semestre

Se participó en la Universidad Tecnológica de Emiliano Zapata (UTEZ), presentando el póster titulado “Sistema de reconocimiento de voz humana y sintética” el día 10 de octubre de 2023 Figura 6.1

- Segundo semestre

Se participó impartiendo una ponencia titulada “Sistema de reconocimiento de voz humana y sintética” en el Instituto Tecnológico de Chilpancingo el día 19 de abril de 2024 Figura 6.2.

- Cuarto semestre

Se presentaron dos artículos. El primero fue expuesto en el Congreso COMIA el 15 de mayo de 2025, con el título “Sistema de reconocimiento de voz humana y sintética” Figura 6.3 y 6.4. El segundo fue presentado en el congreso CITCA el 23 de mayo de 2025, titulado “*Review of AI-generated Voice Cloning Tools*” Figura 6.5 y 6.6.



Figura 6.1 Participación de primer semestre.



EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA



**EL TECNOLÓGICO NACIONAL DE MÉXICO
A TRAVÉS DEL INSTITUTO TECNOLÓGICO DE CHILPANCINGO
OTORGAN EL PRESENTE**

RECONOCIMIENTO

A

JIBRÁN ZAEDT RODRÍGUEZ GARCÍA

POR HABER IMPARTIDO LA PONENCIA **"SISTEMA DE
RECONOCIMIENTO DE VOZ HUMANA Y SINTÉTICA"**,
DESARROLLADA COMO PARTE DE LAS ACTIVIDADES DEL
COLOQUIO DE INTELIGENCIA ARTIFICIAL 2024.

CHILPANCINGO DE LOS BRAVO, CRO., A 9 DE ABRIL DE 2024.

ARELI BARCENAS NAVA

DIRECTORA



Figura 6.2 Participación en segundo semestre.

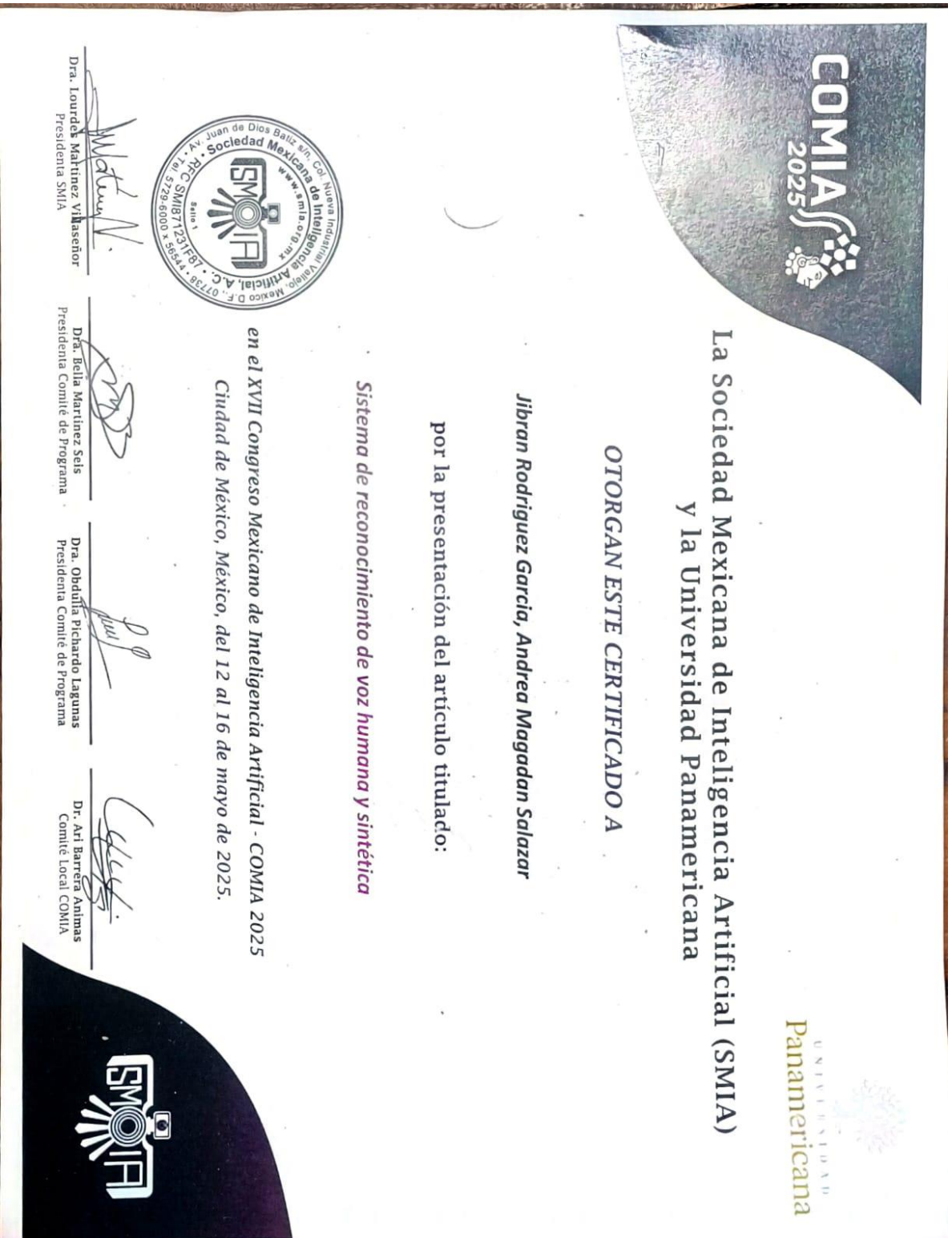


Figura 6.3 Artículo presentado en COMIA.

Sistema de reconocimiento de voz humana y sintética

Jibran Zaedt Rodríguez García, Andrea Magadán Salazar

TecNM/CENIDET

Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET),
México

{m23ce077, andrea.ms}@cenidet.tecnm.mx

Resumen. El advenimiento de la clonación de voz por inteligencia artificial ha revolucionado el campo de la síntesis de voz, ofreciendo una autenticidad y personalización sin precedentes. Las aplicaciones de esta tecnología son numerosas y diversas, abarcando el sector del entretenimiento, la accesibilidad, el marketing digital y enfoques pioneros para la creación de contenido y la comunicación digital. Este artículo analiza algunas de las principales plataformas de clonación de voz y sus aplicaciones prácticas. Examina la funcionalidad, características e impacto de cada plataforma en la comunicación. Luego, la discusión aborda las consideraciones éticas clave que rodean su uso, incluyendo el fraude, la privacidad, la desinformación y el impacto potencial en el empleo de los artistas de voz. El potencial de la clonación de voz para ofrecer experiencias completamente nuevas en muchas áreas es significativo, incluido el uso de voces de personajes en películas y la creación de audiolibros con voces históricas. Además, también proporciona información sobre los motores y algoritmos subyacentes que impulsan estas aplicaciones. También explica cómo la integración de redes neuronales con modelos de alto nivel permite la personalización de voces digitales. Finalmente, el artículo discute la necesidad de un marco ético y regulatorio que garantice prácticas responsables de clonación de voz que protejan los derechos individuales y el valor del trabajo humano en el entorno tecnológico en evolución.

Palabras clave: Herramientas de clonación de voz, algoritmos de clonación, aplicaciones de clonación, inteligencia artificial.

Human and Synthetic Speech Recognition System

Abstract. The advent of artificial intelligence voice cloning has revolutionized the field of speech synthesis, offering unparalleled authenticity and personalization. The applications of this technology are numerous and diverse applications, spanning the entertainment sector, accessibility, digital marketing, and pioneering approaches to content creation and digital communication. This paper analyses some of the main voice cloning platforms and their practical applications. It examines the functionality, features, and impact of

Figura 6.4 Primera página del artículo presentado en COMIA.



Educación
Secretaría de Educación Pública



TECNOLÓGICO
NACIONAL DE MÉXICO

EL TECNOLÓGICO NACIONAL DE MÉXICO
A TRAVÉS DEL CENTRO NACIONAL DE INVESTIGACIÓN
Y DESARROLLO TECNOLÓGICO

OTORGA EL PRESENTE

RECONOCIMIENTO

A

JIBRAN ZAEDT RODRIGUEZ GARCIA | ANDREA MAGADÁN SALAZAR |
NIMROD GONZALEZ FRANCO

Por su destacada participación con la Contribución: ARTÍCULO PARA REVISTA, en el
área: CIENCIAS DE LA COMPUTACIÓN, con el Título:
REVISIÓN DE HERRAMIENTAS DE CLONACIÓN DE VOZ MEDIANTE INTELIGENCIA
ARTIFICIAL

presentada en el marco del 4.º Congreso Internacional de Tecnología y Ciencia
Aplicada (CITCA), celebrado en el Centro Nacional de Investigación y Desarrollo
Tecnológico del 21 al 23 de mayo de 2025.

Cuernavaca, Morelos, 23 de mayo de 2025

LORENZO O. HERNÁNDEZ
Director del CENIDET

MANUEL ADAM MEDINA
Coordinador General
4.º CITCA



Sello Digital:

aGYCgFZKo4M91pn7qocDfQdVqsls/XONuQZ1/R4azGp5iK9SkOdstK7gt+qVmsfQprT1smkY
jtfZBB3UAA4orBuubMgWqG1/Peg/PknWF26QEVwoPHYERhwVYRSIXFKPD+XNgT+WbdgywL
2e1te4qyJ43tjV3d3fW+pW13seXRSReudBJftYHyXCUNSIeM9RoJHbyGBaZ/1BKCV13KU1YV
OOUWaiIn/MDeebfP1Tt1P+L1SE8RjxL78t7pnzoa/Xh1gnsWq348H179oXRPPTPOu6t4cbEV1
4N1XwJlms7/nlaRJoCuRSPOnb97nxCBqSvZ1WfIxhL8cXL/q8qFHA==

CITCA225A25 <http://constancias.cenidet.tecnm.mx>

Figura 6.5 Artículo presentado en CITCA.

Review of AI-generated Voice Cloning Tools

Jibran Zaedt Rodríguez García*, Andrea Magadán Salazar*, Nimrod Gonzalez Franco*

*TECNM/Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, Morelos, México
(email: {m23ce077, andrea.ms, nimrod.gf}@cenidet.tecnm.mx).

Abstract: The advent of artificial intelligence voice cloning has revolutionized the field of speech synthesis, offering unparalleled authenticity and personalization. The applications of this technology are numerous and diverse, spanning the entertainment sector, accessibility, digital marketing, and pioneering approaches to content creation and digital communication. The paper analyses some of the main voice cloning platforms and their practical applications. It reviews the functionality, features, and impact of each platform on communication. Furthermore, it also discusses the primary ethical concerns associated with their utilization, including fraud, privacy, misinformation, and the potential impact on the employment of voice artists. Voice cloning offers entirely new experiences in many areas, including the use of character voices in films and the creation of audiobooks with historical voices. Furthermore, it provides insights into the underlying engines and algorithms that power these applications. It also explains how the integration of neural networks with high-level models enables the customization of digital voices. The paper concludes by discussing the necessity for an ethical and regulatory framework that ensures responsible voice cloning practices, protecting individual rights and the value of human labor in the evolving technological environment.

Palabras clave: Voice cloning tools, cloning algorithms, cloning applications, artificial intelligence.

1. INTRODUCCIÓN

En la actualidad, los asistentes de voz o altavoces inteligentes están ganando protagonismo y ya forman parte de nuestra vida. Por ejemplo, se pueden controlar otros dispositivos como termostatos, aires acondicionados, luces, refrigeradores, televisores, entre otros, mediante la voz. Los teléfonos inteligentes han sido los precursores en estos desarrollos de la voz como elemento primordial. Las aplicaciones son variadas y van desde los asistentes personales para ejecutar órdenes, para reconocer dictados, leer textos, para añadir realismos en los videojuegos, etc.

Los sistemas de reconocimiento de voz y generación de voz que pueden procesar de forma natural un diálogo entre humano y máquina se encuentran bajo el término de sistemas o interfaces de voz natural que se destinan a aplicaciones de cliente-servidor en entornos conversacionales. La generación de voz es el proceso mediante el cual un dispositivo inteligente produce secuencias de habla artificial. El reconocimiento de voz es el proceso mediante el cual las computadoras interpretan y digitalizan las señales de voz, tanto para su análisis de contenido como para la interpretación de órdenes.

El proceso de conversión de texto a voz en los asistentes virtuales consta de tres etapas principales (Achanta et al., 2021).

1. *Entrada de texto y conversión fonémica:* El texto se transforma en una cadena de fonemas, incluyendo puntuación y límites de palabras. Esto permite al modelo capturar mejor la prosodia y los ritmos del habla.

2. *Creación de espectrogramas Mel con Tacotron:* Los fonemas se convierten en espectrogramas Mel utilizando una red basada en la atención secuencia a secuencia, Tacotron. Esta red emplea el enfoque secuencia a secuencia con capas semejantes a la red Long Short-Term Memory en la secuencia para procesar y generar muchos fotogramas del espectrograma al mismo tiempo, lo que lo hace más eficiente y de alta calidad.

3. *Conversión a audios con WaveRNN:* El espectrograma de Mel alimenta a una red neuronal autorregresiva llamada WaveRNN para generar audio, muestra por muestra. En ese nivel, la señal de audio se genera a partir del espectrograma mismo por la red neuronal, y el control de la velocidad y la calidad se realiza mediante una optimización adicional. Este proceso no utiliza los datos de voz exactos que uno podría desear simular. En cambio, utiliza una voz de archivo en el proceso de creación de los fonemas asociados para el texto dado y, por lo tanto, realmente pierde mucho en términos de fidelidad para simular la voz de una persona real.

En los últimos años ha surgido otra área de desarrollo conocida como clonación de voz. En términos sencillos, la clonación de voz es el proceso de copiar la voz de una persona para reproducirla o generarla en un contexto diferente al original. Es producir una voz artificial que tenga las mismas características (suene igual) como si la hubiera pronunciado una persona objetivo (Achanta et al., 2021).

La clonación de voz no es nueva; sin embargo, las nuevas herramientas de inteligencia artificial logran mayores niveles de autenticidad y personalización. Este progreso se ha utilizado para replicar la voz humana de formas que ensalzan

Referencias

- [1] C. Canton Ferrer, B. Dolhasky, B. Pflaum, J. Bitton, J. Pan, and J. Lu, "Deepfake Detection Challenge Results: An open initiative to advance AI." Accessed: Oct. 25, 2023. [Online]. Available: <https://ai.meta.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/>
- [2] "Reset." Accessed: Oct. 19, 2023. [Online]. Available: <https://www.reset.tech/>
- [3] P. Marcek, "POZOR, POZOR, POZOR! 😬😬😬 | Facebook," Facebook. Accessed: Oct. 19, 2023. [Online]. Available: <https://www.facebook.com/peter.marcek.5/posts/276572018586296/>
- [4] M. Uhrik, "Aj Michal už dostal rozum 😊... | Facebook," Facebook. Accessed: Oct. 19, 2023. [Online]. Available: <https://www.facebook.com/ing.milan.uhrik/posts/pfbid02gmCjGeQ2PZZuFv7T7HMGG6yVe7WvyqfL1n6Me5SJXMHZc9Ww76yKWG2zeYUhd9S7I>
- [5] R. Barca, "Údajná nahrávka telefonátu predsedu PS a novinárky Denníka N vykazuje podľa expertov početné známky manipulácie | Fakty," *AFP farkty*, Sep. 28, 2023. Accessed: Oct. 19, 2023. [Online]. Available: <https://fakty.afp.com/doc.afp.com.33WY9LF>
- [6] P. Verma and will Oremus, "AI voice clones are all over social media, and they're hard to detect - The Washington Post," *The Washington Post*, Oct. 13, 2023. Accessed: Oct. 17, 2023. [Online]. Available: <https://www.washingtonpost.com/technology/2023/10/13/ai-voice-cloning-deepfakes/>
- [7] "GitHub: Let's build from here · GitHub." Accessed: Jun. 12, 2024. [Online]. Available: <https://github.com/>
- [8] "Discord - Un chat de grupo lleno de juegos y diversión." Accessed: Jun. 12, 2024. [Online]. Available: <https://discord.com/>
- [9] "GitHub - IAHispano/Aplio-RVC-Fork: Enhanced RVC Variant: Optimized Performance Through Modifications, Built upon Mangio-RVC-Fork." Accessed: Nov. 30, 2023. [Online]. Available: <https://github.com/IAHispano/Aplio-RVC-Fork>
- [10] "GitHub - RVC-Project/Retrieval-based-Voice-Conversion-WebUI: Voice data <= 10 mins can also be used to train a good VC model!" Accessed: Nov. 23, 2023. [Online]. Available: <https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI>
- [11] "Free Real Time Voice Changer for PC - Voice.ai." Accessed: Nov. 23, 2023. [Online]. Available: <https://voice.ai/>
- [12] "AI Voice Generator & Text to Speech | ElevenLabs." Accessed: Apr. 29, 2024. [Online]. Available: <https://elevenlabs.io/>
- [13] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," 1995. [Online]. Available: <https://www.researchgate.net/publication/240312210>
- [14] "GitHub - Pradeepiit/hf0: Hybrid f0 estimation using Convolutional Neural Network." Accessed: May 27, 2024. [Online]. Available: <https://github.com/Pradeepiit/hf0>
- [15] E. Díaz, "Sistema de reconocimiento de género mediante el análisis acústico de voz," Centro Nacional de Investigación y Desarrollo Tecnológico, 2022.
- [16] N. M. Müller, K. Pizzi, and J. Williams, "Human Perception of Audio Deepfakes," in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, New York, NY, USA: ACM, Oct. 2022, pp. 85–91. doi: 10.1145/3552466.3556531.
- [17] S.-Y. Lim, D.-K. Chae, and S.-C. Lee, "Detecting Deepfake Voice Using Explainable Deep Learning Techniques," *Applied Sciences*, vol. 12, no. 3926, Apr. 2022, doi: 10.3390/app12083926.
- [18] X. Li *et al.*, "Replay and Synthetic Speech Detection with Res2net Architecture," Oct. 2020.
- [19] L. Blue *et al.*, "Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction."
- [20] Y. Bensoussan, J. Pinto, M. Crowson, P. R. Walden, F. Rudzicz, and M. Johns, "Deep Learning for Voice Gender Identification: Proof-of-concept for <scp>Gender-Affirming</scp> Voice Care,"

- Laryngoscope*, vol. 131, no. 5, May 2021, doi: 10.1002/lary.29281.
- [21] D. Kwasny and D. Hemmerling, "Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks," *Sensors*, vol. 21, no. 14, p. 4785, Jul. 2021, doi: 10.3390/s21144785.
 - [22] R. Mohammad Sahar, T. Srivinas Rao, S. Anuradha, and B. Srinivasa Rao, "Performance Analysis of ML Algorithms to Detect Gender Based on Voice," 2021. doi: 10.3233/APC210192.
 - [23] V G Nandan, Sukruth Shivakumar, J Sangeetha, Mukund Pandurang Nayak, and Nishanth S K, "A Comparative Study of Deep Learning and Machine Learning Approaches in Speech Emotion and Gender Recognition System," 2021.
 - [24] M. M. Nasef, A. M. Sauber, and M. M. Nabil, "Voice gender recognition under unconstrained environments using self-attention," *Applied Acoustics*, vol. 175, p. 107823, Apr. 2021, doi: 10.1016/j.apacoust.2020.107823.
 - [25] S. Chamishka *et al.*, "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling," *Multimed. Tools Appl.*, vol. 81, no. 24, pp. 35173–35194, Oct. 2022, doi: 10.1007/s11042-022-13363-4.
 - [26] A. Davletcharova, S. Sugathan, B. Abraham, and A. P. James, "Detection and Analysis of Emotion from Speech Signals," *Procedia Comput. Sci.*, vol. 58, pp. 91–96, 2015, doi: 10.1016/j.procs.2015.08.032.
 - [27] A. Lausen and K. Hammerschmidt, "Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters," *Humanit. Soc. Sci. Commun.*, vol. 7, no. 1, p. 2, Jun. 2020, doi: 10.1057/s41599-020-0499-z.
 - [28] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual Cues in Nonverbal Vocal Expressions of Emotion," *Quarterly Journal of Experimental Psychology*, vol. 63, no. 11, pp. 2251–2272, Nov. 2010, doi: 10.1080/17470211003721642.
 - [29] V. M. Reddy, T. Vaishnavi, and K. P. Kumar, "Speech-to-Text and Text-to-Speech Recognition Using Deep Learning," in *Proceedings of the 2nd International Conference on Edge Computing and Applications, ICECAA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 657–666. doi: 10.1109/ICECAA58104.2023.10212222.
 - [30] X. Zhou, Z.-H. Ling, and S. King, "The Blizzard Challenge 2020," International Speech Communication Association, Oct. 2020, pp. 1–18. doi: 10.21437/vcc_bc.2020-1.
 - [31] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, Feb. 2023, doi: 10.1007/s10489-022-03766-z.
 - [32] Y. Wang *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," Mar. 2017.
 - [33] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," Oct. 2018.
 - [34] S. O. Arik *et al.*, "Deep Voice: Real-time Neural Text-to-Speech," Feb. 2017.
 - [35] S. Arik *et al.*, "Deep Voice 2: Multi-Speaker Neural Text-to-Speech," May 2017.
 - [36] W. Ping *et al.*, "Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning," Oct. 2017.
 - [37] A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio," Sep. 2016.
 - [38] "Kits AI - the ultimate AI voice generation toolkit to supercharge your music." Accessed: Nov. 23, 2023. [Online]. Available: <https://app.kits.ai/>
 - [39] "Voice Clone|Text to song|Text to speech_MyVocal.ai." Accessed: Nov. 23, 2023. [Online]. Available: <https://www.myvocal.ai/>
 - [40] "speechify." Accessed: Nov. 29, 2023. [Online]. Available: https://speechify.com/?via=blue&source=fb-for-mobile&landing_url=https%3A%2F%2Fspeechify.com%2Fblog%2Fhow-to-clone-your-voice-with-ai%2F&gclid=CjwKCAiApaarBhB7EiwAYiMwqklddh5eznuS-71LFjpjYvmh1pnSdOsT7ICR4wclMuasl-bNZJyTUXoCHPQQAvD_BwE
 - [41] "AI Voice Generator with Text to Speech and Speech to Speech." Accessed: Nov. 30, 2023. [Online]. Available: <https://www.resemble.ai/>

- [42] "CereVoice Me Voice Cloning Service | CereProc Text-to-Speech." Accessed: Nov. 30, 2023. [Online]. Available: <https://www.cereproc.com/en/products/cerevoiceme>
- [43] "FakeYou Celebrity Voice Generator | FakeYou." Accessed: Nov. 30, 2023. [Online]. Available: <https://fakeyou.com/>
- [44] B. TORRES GALLARDO, "LA VOZ Y NUESTRO CUERPO: UN ANÁLISIS FUNCIONAL," in *Revista de Investigaciones en Técnica Vocal*, vol. 1 (2013), 2015, pp. 40–58.
- [45] H. Ullah Sheikh, "WHO IS SPEAKING? MALE OR FEMALE," UNIVERSITY OF MANCHESTER, 2013.
- [46] X. Huang, A. Acero, and H. hon Hon, *Spoken language processing*. 2001.
- [47] C. Rincón and R. Barra, "Diseño, implementación y evaluación de técnicas de identificación de emociones a través de la voz," 2007.
- [48] B. Yang, "The fundamental frequency (f0) distribution of American speakers in a spontaneous speech corpus*," *Phonetics and Speech Sciences*, vol. 16, no. 1, pp. 11–16, Mar. 2024, doi: 10.13064/KSSS.2024.16.1.011.
- [49] Y. S. TAŞPINAR, M. M. SARITAŞ, İ. ÇINAR, and M. KOKLU, "Gender Determination Using Voice Data," *International Journal of Applied Mathematics Electronics and Computers*, vol. 8, no. 4, Dec. 2020, doi: 10.18100/ijamec.809476.
- [50] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A Convolutional Representation for Pitch Estimation," Feb. 2018.
- [51] J. K. MacCallum, A. E. Olszewski, Y. Zhang, and J. J. Jiang, "Effects of Low-Pass Filtering on Acoustic Analysis of Voice," *Journal of Voice*, vol. 25, no. 1, pp. 15–20, Jan. 2011, doi: 10.1016/j.jvoice.2009.08.004.
- [52] R. Bhagat and R. Kaur, "Improved Audio Filtering Using Extended High Pass Filters," *International Journal of Engineering Research & Technology*, vol. 2, no. 6, Jun. 2013.
- [53] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Jun. 2016, doi: 10.1145/2939672.2939785.
- [54] E. J. Carmona, "Tutorial sobre Máquinas de Vectores Soporte (SVM)." [Online]. Available: <https://www.researchgate.net/publication/263817587>
- [55] X. Riley and S. Dixon, "CREPE Notes: A new method for segmenting pitch contours into discrete notes," Nov. 2023, doi: 10.5281/zenodo.8136568.
- [56] "Common Voice." Accessed: Apr. 29, 2024. [Online]. Available: <https://commonvoice.mozilla.org/en/datasets>
- [57] X. Wang *et al.*, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," Nov. 2019.
- [58] victor, "High-pass, Low-pass and Band-pass Filtering Program objectives Restitution of the signal without any filtering (option 0)."
- [59] "GitHub - Pradeepiit/hf0: Hybrid f0 estimation using Convolutional Neural Network." Accessed: Jun. 02, 2024. [Online]. Available: <https://github.com/Pradeepiit/hf0>
- [60] J. Z. Rodríguez García, A. Magadán Salazar, and N. Gonzalez Franco, "Review of AI-generated Voice Cloning Tools," *Tecnología y Ciencia Aplicada*, vol. 8, no. 1, pp. 599–604, 2025, Accessed: Dec. 04, 2025. [Online]. Available: <https://jcyta.cenidet.tecnm.mx/>
- [61] J. Z. Rodríguez García and A. Magadán Salazar, "Sistema de reconocimiento de voz humana y sintética," *Research in Computing Science*, vol. 154, no. 7, pp. 5–17, Oct. 2025, Accessed: Dec. 04, 2025. [Online]. Available: <https://www.rcs.cic.ipn.mx/>