EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Análisis Acústico de Voz para la Identificación de Sexo y Categorización de Edad, en Múltiples Idiomas y Bajo Ambientes No Controlados

presentada por

Lic. Enrique Díaz Ocampo

como requisito para la obtención del grado de

Maestra en Ciencias de la Computación

Director de tesis

Dra. Andrea Magadán Salazar

Cuernavaca, Morelos, México. Diciembre de 2023.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., 27/noviembre/2023

OFICIO No. DCC/203/2023
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 ENRIQUE DÍAZ OCAMPO con número de control M22CE002, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"ANÁLISIS ACÚSTICO DE VOZ PARA LA IDENTIFICACIÓN DE SEXO Y CATEGORIZACIÓN DE EDAD, EN MÚLTIPLES IDIOMAS Y BAJO AMBIENTES NO CONTROLADOS"** 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.

ANDREA MAGADÁN SALAZAR
Directora de tesis

RAÚL PINTO ELÍAS
Revisor 1

MÁXIMO LÓPEZ SÁNCHEZ
Revisor 2

EDUCACIÓN
05 DIC 2023
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
RECIBIDO

SEP TecNM CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
05 DIC 2023
SUBDIRECCIÓN ACADÉMICA

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

Cuernavaca, Mor., 11/diciembre/2023
No. De Oficio: SAC/195/2023
Asunto: Autorización de impresión de tesis

ENRIQUE DÍAZ OCAMPO
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 **"ANÁLISIS ACÚSTICO DE VOZ PARA LA IDENTIFICACIÓN DE SEXO Y CATEGORIZACIÓN DE EDAD, EN MÚLTIPLES IDIOMAS Y BAJO AMBIENTES NO CONTROLADOS"**, 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.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y tecnología al servicio de México"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

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

CMAZ/lmz

Dedicatoria

Para mi esposa, quien siempre ha estado conmigo aplacando los demonios en mi cabeza.

Agradecimientos

- Al Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCYT) por el apoyo económico que me brindó para realizar mis estudios de maestría.
- Al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) por brindarme la oportunidad de superarme profesionalmente al formar parte del programa de Maestría en Ciencias de la Computación.
- A mi directora de tesis, la Dra. Andrea Magadán Salazar, por sus enseñanzas y apoyo brindado.
- A mis revisores, el Dr. Raúl Pinto Elías y el Dr. Máximo López Sánchez, por sus enseñanzas, su tiempo, dedicación, sus correcciones y observaciones durante el desarrollo de la presente tesis.
- A mis profesores, el M.C. José Luis Alcántara y el M.C. David Luviano por sus enseñanzas y asesorías brindadas.
- A mi esposa, Areli Karina Martínez Tapia (Tapoa). Por nunca dejar de apoyarme en tantas noches de desvelo.
- A mis padres, Concepción y Braulio, a mis hermanos Teresa, Erick y Miguel.
- A mis suegros, María Reyna y Felipe de Jesús por su apoyo y cariño.
- A mi padrino, Felipe de Jesús por sus explicaciones del modelo físico.
- A mi sobrinas, Luna Michel y Alondra Mayte, quienes me escuchaban mis charlas sobre Inteligencia Artificial.
- Y a todos mis compañeros de Maestría y a los que cursaban el Doctorado. Gracias por tantas charlas tan interesantes sobre ciencia.

Resumen

El reconocimiento de género (género binario o también conocido como sexo) por voz mediante algún sistema es una actividad diaria que se ejecuta en diversos asistentes personales: Siri, Alexa, Google, etc. Ha sido un problema estudiado principalmente con nativos del idioma inglés, lo cual ocasiona un sesgo cuando el hablante no es nativo o hablante de dicho idioma. Los sistemas propuestos para el reconocimiento de género, se enfocan en la implementación de características profundas, que requiere costo computacional y cuyos resultados solo son entendibles por una computadora, dificultando la interpretabilidad del sistema. En la investigación aquí reportada, se partió de que el reconocimiento de género por voz, puede abordarse mediante el estudio de características robustas extraíbles a partir de un audio que puedan reconocer el género a pesar de no hablar inglés. Se abordó el reconocimiento del género a partir de la voz, mediante características biológicas i.e. Frecuencia fundamental, Intensidad de la Voz, Longitud del Tracto vocal, y Coeficientes Cepstrales de Frecuencias de Mel, en los idiomas: inglés, francés, alemán, chino, español, y thai. El conjunto de datos de voces, se obtuvieron del conjunto de voces *Mozilla Common Voice*. Se implementaron un total de seis metodologías: tres para el reconocimiento de género, y tres para el reconocimiento de género y edad. Con éstas metodologías se obtuvieron valores de reconocimiento de género superiores al 90 %. Los hallazgos indican que los cuartiles de la Frecuencia fundamental, aunado a la estimación del tracto vocal, muestran robustez, cuando se intenta reconocer el idioma inglés y el español. Se concluye que el reconocimiento de género mediante la voz, es factible, sin la necesidad de emplear características de índole profundas, para ello se requiere el estudio estadístico de las características de cada idioma.

Palabras clave: género, frecuencia fundamental, idiomas, sistema, redes neuronales.

Abstract

Recognition of gender (binary gender or also known as sex) by voice through some system is a daily activity that is executed in several personal assistants: Siri, Alexa, Google, etc. It has been a problem studied mainly with native English speakers, which causes a bias when the speaker is not a native or native English speaker. The systems proposed for gender recognition focus on the implementation of deep features, which require computational cost and whose results are only understandable by a computer, making the interpretability of the system difficult. In the research reported here, it was assumed that speech-based gender recognition can be addressed by studying robust features extractable from audio that can recognize gender despite not speaking English. Gender recognition from voice was approached using biological features i.e. Fundamental Frequency, Voice Intensity, Vocal Tract Length, and Mel Frequency Ceptral Coefficients in English, French, German, Chinese, Spanish, and Thai languages. The voice dataset was obtained from the Mozilla Common Voice dataset. A total of six methodologies were implemented: three for gender recognition, and three for gender and age recognition. With these methodologies, gender recognition values above 90 % were obtained. The findings indicate that the quartiles of the Fundamental Frequency, together with the estimation of the vocal tract, show robustness when trying to recognize English and Spanish. It is concluded that gender recognition by means of voice is feasible, without the need to use deep characteristics, for which the statistical study of the characteristics of each language is required.
Keywords: gender, fundamental frequency, languages, system, neural networks.

Contenido

Lista de Tablas

Lista de Figuras

Lista de acrónimos

AT: *Acoustic Treatment* (Tratamiento Acústico).
ACT: *Acoustic and Cepstral Treatment* (Tratamiento Acústico y Cepstral).
AKM: *Acoustic Knowledge Modeling* (Modelado de conocimiento acústico).
MFCC: *Mel Frequency Cepstral Coefficient* (Coeficientes Cepstrales de Frecuencias de Mel.
ROC: *Receiver Operating Characteristic* (Característica Operativa del Receptor).
VTL: *Vocal Tract Length* (Longitud del Tracto Vocal).

1 | Presentación

Dentro de las múltiples ramas del procesamiento del habla, el reconocimiento de características humanas a partir de voz es una de las más populares en los recientes años, porque aplicaciones en la medicina [1, 2], en los centros de llamadas [3, 4], y en la mejoría del desempeño de sistemas enfocados en las interacciones humano-computadora [5].

1.1. Descripción del Problema

De las diversas características existentes en la voz (edad, emoción, acento, etc.), el género es una de las de mayor impacto actual. Esto se debe a que posee cierta información sobre actividades sociales de la persona. El género puede concebirse bajo dos enfoques [6]:

- Enfoque esencialista del género: Existen rasgos biológicos que ayudan a distinguir a un género de otro. La palabra género se usa cómo sinónimo al sexo.
- Enfoque constructivista del género: El género es una construcción social. Está asociado a ciertas expectativas, condicionantes, y costumbres del nicho social al que pertenece la persona.

El reconocimiento de género por voz y los sistemas enfocados en esta tarea, utilizan únicamente el enfoque esencialista del género, debido a que existen múltiples estudios que han demostrado las diferencias acústicas entre hombres y mujeres. Desde los primeros trabajos estadísticos del área, realizados en la década de 1990 (véase [7, 8]), seguido de enfoques con características acústicas y espectrales (véase [9, 10]), hasta los trabajos más recientes utilizando modelos de aprendizaje profundo (véase [4, 11, 12]). No obstante, aún con las múltiples técnicas desarrolladas actualmente, todavía es un desafío reconocer el género de una persona por voz. La razón está relacionada con diversos factores por ejemplo:

- **Relacionados con el hablante**:
  - Fisiológicos: Diferencias de longitud y grosor de las cuerdas vocales, la longitud del tracto vocal, la salud y edad de la persona, etc.
  - Anímicos: Cambios de estado emocional de la persona.
  - Lingüísticos: Idiomas conocidos por el hablante y su acento en cada uno de ellos.
  - Estilo del habla: Si el hablante está cantando, hablando, susurrando o solo emite sonidos por la boca (e.g. silbando).
- **Relacionados con el ambiente de grabación**:
  - Si la grabación se realiza en entornos clínicos o controlados.
  - Si la grabación se realiza en entornos con múltiples fuentes de ruido (no controlados).
- **Relacionados con el equipo de grabación y la grabación**
  - Duración del audio.
  - Calidad del equipo de grabación.

En este proyecto se enfoca en la detección del género mediante la voz. Por lo anterior, se requieren análisis estadísticos que distingan las voces de hombres y mujeres, pero que muestren robustez ante los diversos factores mencionados.

1.2. Planteamiento del Problema

El problema a resolver es clasificar el género masculino o femenino de una voz adulta, a partir de las características acústicas y cepstrales de una muestra de audio provenientes de ambientes no controlados (véase figura 1.1). Siendo que la mayoría de los sistemas de reconocimiento de género reportados en este documento trabajan con un idioma que generalmente es el inglés. Generando un sesgo cuando el hablante no es nativo de

dicho idioma. Por otro lado, los conjuntos de voces gratuitos (véase *Mozilla Common Voice* [13]) contienen voces en otros idiomas que suelen presentar más voces de un género. Por lo anterior, se requiere verificar que las características utilizadas en la detección de género en un idioma pueden emplearse en diferentes lenguajes sin repercutir en el desempeño del sistema. Además, para efectos de esta investigación, se establece que un audio tomado de un conjunto de datos públicos está en un ambiente no controlado cuando para cada audio varia: el acento, la calidad del audio con respecto al dispositivo de grabación, el ruido ambiental, la duración del audio, la salud y estado emocional de la persona, la frase dicha y la edad.

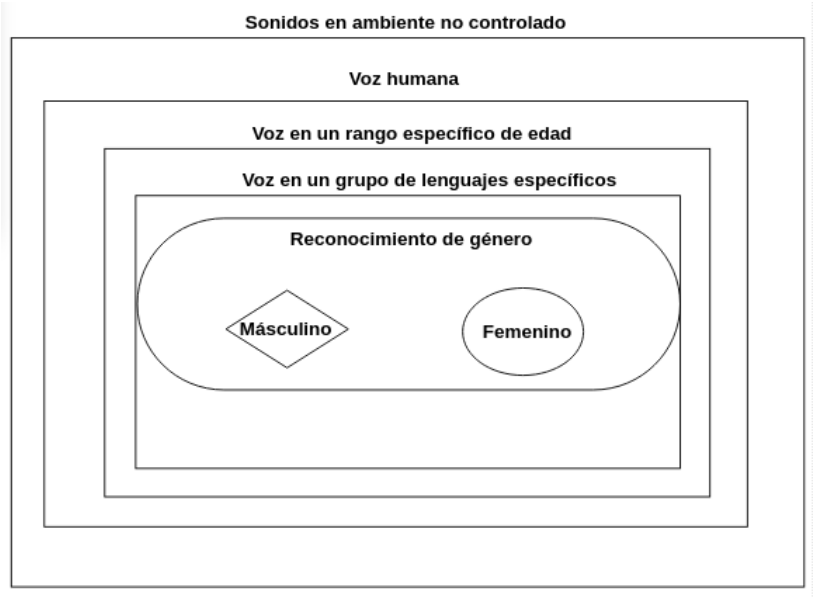


Figura 1.1: Esquema general de la descripción del problema.

1.3. Complejidad del Problema

La clasificación del género mediante la voz esta influenciada por múltiples factores.

- **Biológicos**:
  1. La edad modifica la elasticidad de las cuerdas vocales afectando su vibración y por ende su sonido.
  2. Las condiciones médicas que afecten al tracto vocal y a la laringe alteran el sonido de la voz.
- **Linguísticos**:
  1. Las personas tienen diferentes acentos y pronunciaciones de las mismas palabras por lo que aún cuando su voz es grabada en un ambiente controlado con un texto específico se dificulta su clasificación.
  2. Las personas tienen diferentes velocidades de habla, lo que dificulta tomar una muestra uniforme en ambientes no controlados para obtener información relevante de su género.
- **Acústicos**:
  1. La calidad de la grabación de la voz está sujeta a equipos sofisticados.
  2. Los ruidos de ambiente limitan la precisión del clasificador.
- **Computacionales**:
  1. El muestreo y cuantización de una señal de voz están sujetos al tipo de equipo que los está almacenando por lo que su información se ve comprometida.
  2. Los preprocesamientos del audio requieren de cierto conocimiento técnico para
     - Dividir la señal en diferentes marcos de tamaños adecuados según el archivo sin perder información.

- Discernir que filtro es el más adecuado para las señales usadas.
- Discernir que tipo de algoritmo se utilizará para calcular los parámetros acústicos de la señal.
- Transformar las señales de amplitud en señales de frecuencia mediante funciones complejas, por ejemplo, la transformada de Fourier. De este modo poder visualizar el sonido mediante espectros y tener un tipo de análisis más sofisticado.

1.4. Propuesta de Solución

El objetivo general, los objetivos específicos, los alcances y las limitaciones se muestran a continuación. Además, se presentan las metodologías propuestas.

1.4.1. Objetivo General

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 grabadas en ambientes no controlados.

1.4.2. Objetivos Específicos

- Revisar el estado del arte de los sistemas de reconocimiento de género por voz.
- Realizar una búsqueda de bases de datos públicas con diversos audios de voz categorizados por género e idioma.
- Realizar las diversas etapas de preprocesamiento de la voz para atenuar el ruido de la señal.
- Realizar la extracción de características que permitan determinar el género de la voz.
- Realizar la selección de las características más relevantes para el problema de clasificación.
- Implementar el clasificador y sus métricas acordes al problema para la detección de género.
- Comparar los resultados obtenidos con los reportados en el estado del arte.

1.4.3. Alcances

- El sistema realiza el reconocimiento del género en las categorías masculino y femenino en al menos dos lenguajes (español e inglés) mediante un archivo de audio en formato MP3.
- No se realiza el reconocimiento de características paralingüística, i.e. el estado anímico, acento, etc.
- Los audios pasan por un preprocesamiento para atenuar la mayor cantidad de ruido posible.
- El sistema realiza la detección del género de una persona mediante un archivo de audio en formato MP3.
- Las voces son obtenidas a partir del conjunto de datos de *Mozilla Common Voice*.
- Las características obtenidas de la voz son las obtenidas por el lenguaje de programación Python.
- El sistema hace uso de las métricas precisión, exhaustividad, exactitud y puntaje F1.
- La cantidad de audios a estudiar es una muestra del total del conjunto de datos, a manera de preservar cierta regularidad de los individuos a estudiar.

1.4.4. Limitaciones

- El audio a analizar sólo debe contener a un solo hablante y sin cambios de lenguaje a lo largo de la grabación.
- El sistema no realiza el reconocimiento de las acciones solicitadas por él o la hablante.
- Los audios que presentan una cantidad de ruido mayor a un valor esperado no se se consideran.
- Se consideraron los audios de personas dentro del rango de edad de 20 hasta los 65 años.
- El diseño del clasificador está sujeto a las características del equipo computacional disponibles.

1.5. Metodologías de Solución

Inicialmente, la propuesta de solución consistía en el diseño, entrenamiento y validación de un sistema de reconocimiento de género mediante archivos de audio de voz en ambientes no controlados. Sin embargo, gracias a las múltiples colaboraciones con el grupo interdisciplinario *Voice Collab AI*[1] dirigido por la Dra. Yael Bensoussan, aunado a los diversos contactos realizados con investigadores nacionales e internacionales, se logró expandir los alcances establecidos. Por lo tanto, se han realizado un total de seis metodologías, tres de ellas enfocadas en el reconocimiento de género y las restantes en el reconocimiento de género y edad. A través de estas metodologías, se ha podido obtener información detallada sobre los patrones de género y edad en los datos analizados. La diversidad de enfoques permitió explorar diferentes posibilidades y evaluar la efectividad de cada método en función de su propósito específico. Para poder entrar en detalle en cada una de ellas, se brinda una introducción a la Teoría Fuente-Filtro que modela la voz de tal forma que la considera una señal analógica en el marco teórico. Sin embargo, en términos generales se tienen tres tipos de sistemas:

1. Un primer sistema con detección de género mediante audios con sus respectivas variantes asociadas a diferentes características acústicas o cepstrales extraídas.
2. El segundo sistema con detección de voces de adolescentes y adultas (en las categorías adolescente-masculino, adolescente-femenino, adulto y adulta) y detección de rango de edad en décadas (en las categorías menos de 19 años, 20 a 29, 30 a 39, 40 a 49, 50 a 59 y más de 60) .
3. El tercer sistema con un módulo de detección de género y rango de edad (en doce categorías del tipo género-edad, e.g. hombre de 20 a 29 años).

Las metodologías y las arquitecturas de los sistemas son discutidas en las siguientes secciones. Los esquemas generales de las propuestas de solución planteadas pueden verse en las Figuras 1.2 y 1.3. En el siguiente listado se hace mención de sus características principales.

1.5.1. Metodologías de Reconocimiento de Género

1. **Primera metodología (véase Fig. 1.2 inciso A):**

   El objetivo fue contar con un sistema entrenado con características tales que:
   - Puedan utilizarse en múltiples idiomas.
   - Robustas ante escenarios con instancias balanceadas y no balanceadas.
   - Robustas ante conjuntos de datos que contengan un audio por hablante o múltiples audios por hablantes.

   En este caso, se propuso el análisis del tono en el rango de 75 a 600 Hertz. Para cada audio se extrajeron 8 características estadísticas derivadas del tono, a saber, mínimo del tono, primer cuartil del tono, mediana del tono, media del tono, tercer cuartil del tono, máximo del tono, desviación estándar del tono y la etiqueta de edad en décadas (veintes, treintas, etc.). Se evaluaron en 5 lenguajes: inglés, español, alemán, francés y chino[2]. Luego, se construyeron dos escenarios, instancias balanceadas e instancias no balanceadas. Cada escenario se enfoca en dos casos, voces únicas (un audio por hablante) y voces repetidas (múltiples audios por hablante). Finalmente, se empleó una validación cruzada de 10 carpetas y las métricas precisión, exhaustividad, puntaje F1, área bajo la curva ROC y exactitud para medir sus prediciones en las categorías de hombre y mujer.

[1] https://www.voicecollab.us

[2] De hecho se usó una mezcla de chino Taiwan, chino Mandarín y chino de Honk Kong.

2. **Segunda metodología (véase Fig. 1.2 inciso B):**

   El objetivo de esta segunda metodología fue evaluar los audios de la primer metodología que no presentan un error de estimación de octava. Esto es, se implementó un filtro que descarta los audios cuya mediana del tono presentaran un valor mayor a 350 Hertz. Los escenarios y las métricas fueron las mismas que en la primer metodología.

3. **Tercera metodología (véase Fig. 1.2 inciso C):**

   El objetivo de la tercera metodología consistió en evaluar el desempeño de un grupo de características para el reconocimiento de género utilizadas en el estado del arte denominadas coeficientes cepstrales de frecuencias de Mel (*MFCC* por sus siglas en inglés). Para cada uno de los audios, se aplica un filtro de detección de voz (i.e. se extrae únicamente las ventanas de este que contengan voz). Posteriormente, se realiza la extracción de la mediana de los *MFCC*. Analizándolos en tres grupos de características: 20, 26 y 30 coeficientes. Finalmente, se empleó una validación cruzada de 10 carpetas y las métricas precisión, exhaustividad, puntaje F1, área bajo la curva ROC y exactitud para medir sus prediciones en las categorias de hombre y mujer. Los idiomas utilizados fueron Español e Inglés en ambientes controlados y no controlados.

1.5.2. Metodologías de Reconocimiento de Género y Edad

Las siguientes tres metodologías consisten en la detección del género y una estimación de la edad del hablante.

1. **Primera Metodología (véase Fig. 1.3 inciso A):**

   El objetivo de esta metodología fue evaluar el desempeño de la estimación del tracto vocal como característica en la detección de voces adolescentes (con 19 años o menos) y voces adultas (de 20 años o más) en ambos géneros y en idiomas tonales y no tonales. Para esto, se empleó un filtro de detección de voz, y se extrajeron dos grupos de características: derivadas del tono en un rango de 75 a 350 Hertz[3] y la estimación tanto de la longitud del tracto vocal y la del cuarto formante. Se construye un vector de características con ambos grupos. Se realiza una validación cruzada de diez carpetas y el empleo de dos tipos de algoritmos: perceptrones multicapa y bosques aleatorios. De este modo, se tienen las predicciones: Hombre de 20 años o más, mujer de veinte años o más, hombre de 19 años o menos, y mujer de 19 años o menos.

2. **Segunda Metodología (véase Fig. 1.3 inciso B):**

   El objetivo de esta segunda metodología fue reconocer sí la voz del hablante pertenece a un adolescente (masculino o femenino de 19 años o menos) o un adulto (hombre o mujer de 20 años o más) y estimar su edad en décadas. Esta metodología está basada en los resultados de la primera agregando los siguientes aspectos:
   - Se implementa un filtro de substracción espectral para la atenuación de ruidos.
   - Se eliminan las ventanas sin voz y se realiza un traslape con las ventanas restantes.
   - Para la detección de género de voces de adolescentes y adultos, se diseñó una red neuronal de perceptrones multicapa denominada **Red de tratamiento acústico** o *Acoustic Treatment*, que empleó las características estadísticas derivadas del tono, la estimación del tracto vocal y el cuarto

[3]Para atenuar el error de octava.

formante.

- Para la estimación de la edad en décadas, se añade un segundo grupo de características que contemplan la mediana de los primeros 30 coeficientes cepstrales, la extracción de características estadísticas de la intensidad de la voz y los primeros cuatro formantes. Además contempla únicamente la mediana del tono extraída en el rango de 75 a 350 Hertz. Posteriormente, se diseñó una segunda red de perceptrones multicapa denominada **Red de tratamiento acústico y Cepstral** o *Acoustic and Cepstral Treatment*, para la detección de edad en décadas: menos de 19 años, 20 a 29, 30 a 39, 40 a 49, 50 a 59 y 60 o más.
- Para la validación de ambas redes, se dividió en conjunto de entrenamiento (80 %), conjunto de prueba (10 %), y conjunto de validación (10 %).

3. **Tercera Metodología (véase Fig. 1.3 inciso C):**
   El objetivo de la tercera metodología fue el reconocimiento del género y la edad en décadas del hablante mediante una sola red neuronal. En cada audio se empleó el mismo preprocesamiento de la metodología anterior. Además, se utilizó una combinación de cuatro grupos de características extraídas:
   - Derivación de estadísticas del tono.
   - Extracción de la mediana de los primeros cuatro formantes.
   - Extracción de la mediana de los primeros 30 *MFCC*.
   - Extracción de estadísticas derivadas de la intensidad de la voz.

   Finalmente, se realizó una tercera red neuronal tipo perceptrón multicapa denominada **Red para modelado de conocimiento acústico** o *Acoustic Knowledge Modeling*. Para la validación de ambas redes, se dividió en conjunto de entrenamiento (80 %), conjunto de prueba (10 %), y conjunto de validación (10 %).

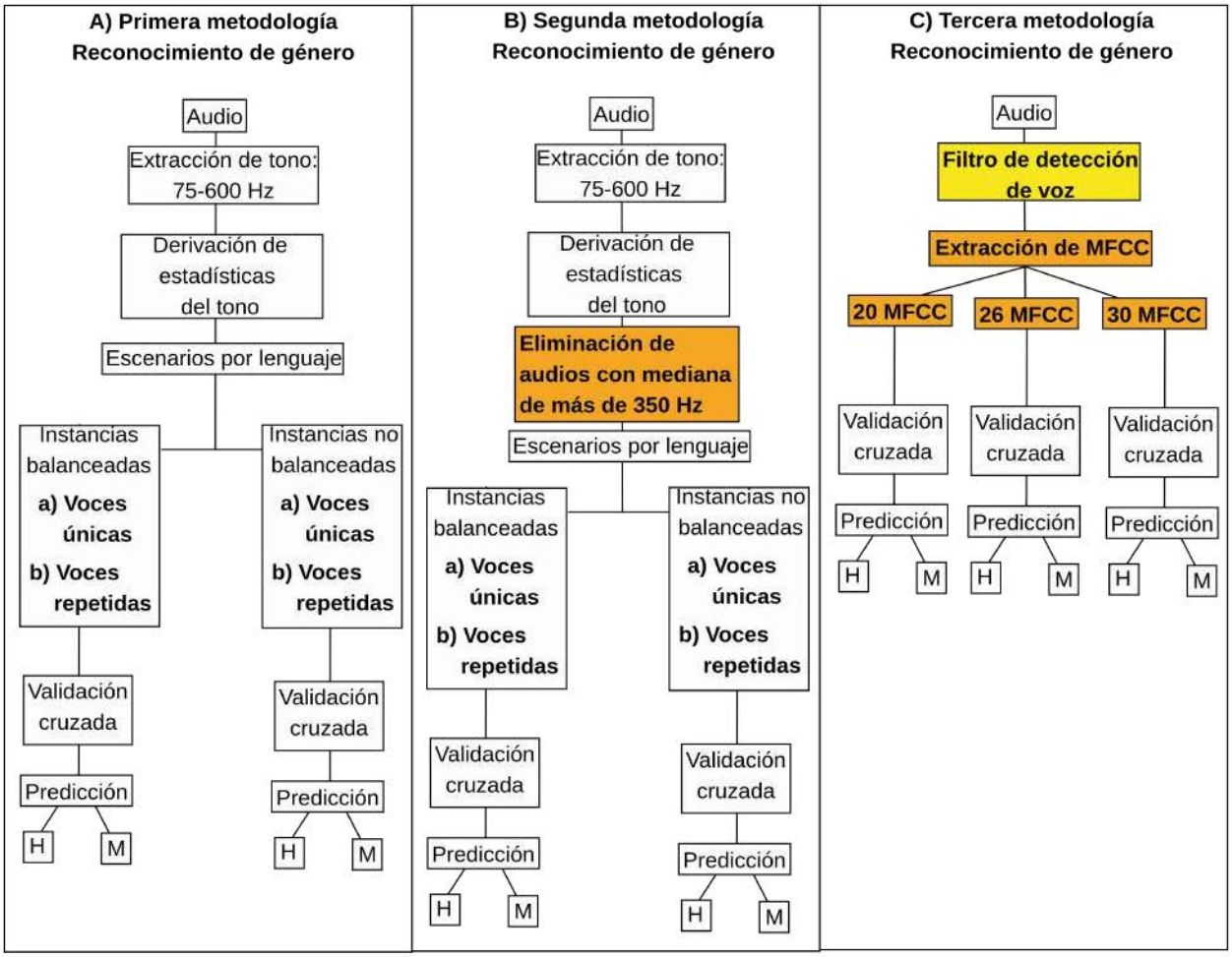


Figura 1.2: Esquema de las metodologías de reconocimiento de género (H y M se utilizan para resumir las palabras hombre y mujer, respectivamente).

1.6. Estructura de la Tesis

La estructura de la tesis se compone de los siguientes capítulos:

- **Capítulo 1: Presentación**. Donde se describe el problema, sus objetivos, alcances, limitaciones, y las metodologías propuestas. Finalmente, se describe la estructura del documento.
- **Capítulo 2: Marco de Referencia**. Donde se detallan los antecedentes, el estado del arte, la propuesta de solución con sus respectivos objetivos, alcances y limitaciones. Además, se detallan las seis metodologías propuestas.
- **Capítulo 3: Marco Teórico**. Donde se brindan las bases biológicas, matemáticas y computacionales empleadas en cada metodología. Además, del estudio de las características empleadas para cada sistema.
- **Capítulo 4: Análisis, Parámetros y Diseño de los Sistemas**. Donde se brinda una exposición sobre el software y librerías empleadas en este trabajo. Además, se detallan los parámetros y sus valores establecidos para la extracción de las características mediante el software PRAAT. Concluyendo con las arquitecturas de las redes propuestas desglosadas por capas.
- **Capítulo 5: Experimentación y Resultados**. Donde se detallan los resultados obtenidos de cada metodología por medio de tablas. Además, se realiza una discusión de los mismos con su respectiva comparación con el estado del arte.

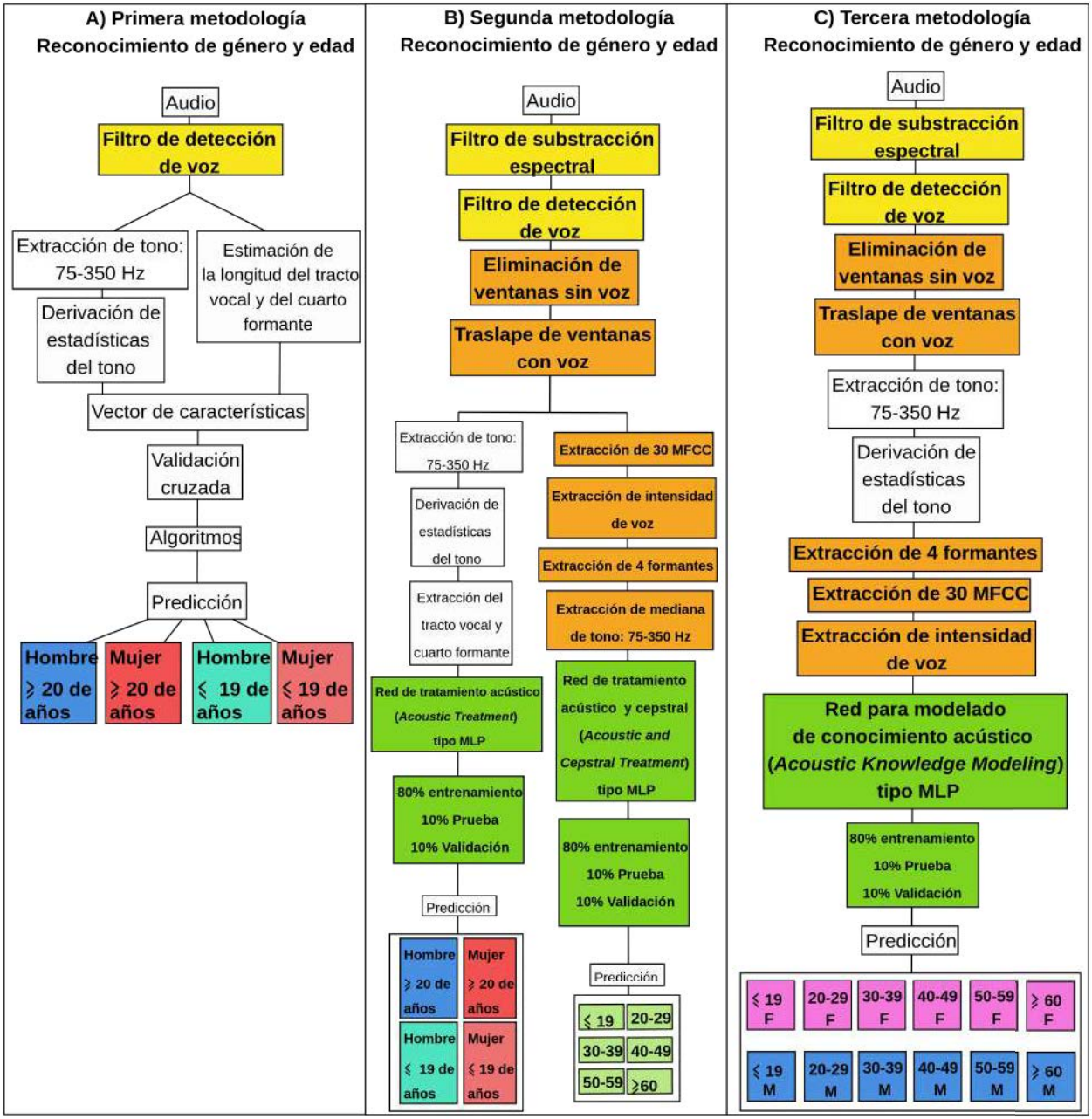


Figura 1.3: Esquema de las metodologías de reconocimiento de género y edad (la letra F denota al género femenino, mientras que M denota al masculino).

- **Capítulo 6: Conclusiones y trabajos futuros**. Se brindan las conclusiones generales y cómo se alcanzaron los objetivos planteados, describiendo además las aportaciones realizadas y el trabajo futuro. Posteriormente, se expone una lista de actividades académicas realizadas durante la maestría.

Finalmente, se adjuntan tres anexos que describen los fundamentos del trabajo de tesis, las actividades académicas y reconocimientos obtenidos con este trabajo de investigación.

2 | Marco de Referencia

2.1. Antecedentes

Para la búsqueda de trabajos anteriores sobre la detección de género mediante análisis acústico de voz en CENIDET, se utilizó el buscador del Repositorio de Tesis. En dicha búsqueda se encontró los siguientes trabajos relacionados al procesamiento de la voz.

Tesis: Ricardo Coronado, "Síntesis de voz para el idioma español usando wavelets"(1999). Tesis de maestría, Tecnológico Nacional de México/CENIDET [14].

- Objetivo: Se desarrolló un sistema de síntesis de voz a partir de una voz grabada, mediante el uso de la herramienta de wavelets.
- Metodología: Se analizaron diferentes audios donde se pronunciaban distintas palabras. Para cada uno de estos, se hizo la separación de las sílabas. Posteriormente, se analizaron los sonidos oclusivos y los fricativos para su extracción de coeficientes de wavelets, tono y envolvente. A partir de estas características, se realizó la generación de voz sintética de cada sílaba detectada. Finalmente se agrupaban todos estos audios para construir la voz sintética.
- Resultados: Los wavelets son una herramienta que requiere menor información para la reconstrucción de una señal de voz. Sin embargo, la calidad de las voces se vió afectada, debido a que el proceso empleado discrimina gran parte del espectro de frecuencia de la señal de voz.

Tesis: Roberto Hernández, "Detección del estado emocional mediante la voz en español de México"(2016). Tesis de maestría, Tecnológico Nacional de México/CENIDET [15].

- **Objetivo**: Se implementó un modelo de clasificación que en conjunto con características acústicas, permitió reconocer la emoción de una voz en un archivo de audio proveniente de dos bases de datos emocionales.
- **Metodología**: La primer base de datos emocional Emo_voz_mx1 consistió en 1541 instancias dividas en las emociones de disgusto, felicidad, ira, miedo, neutro, sorpresa y tristeza. Mientras que la segunda base EmoWisconsin fue de 2040 instancias divididas en indefinido, inseguro, molesto, motivado, nervioso. neutro y seguro. Posteriormente, para cada audio se obtuvieron 45 características divididas en tres grupos: dominio-tiempo, dominio-frecuencia y prosódicas. Luego, se realizaron estudios con diferentes longitudes de ventanas (20, 25, 30, 35 y 40 milisegundos) y funciones de ventanas (Blackman, Hamming y Hanning). Finalmente en cada caso se usaron los clasificadores Perceptrón multicapa, Bayes ingenuo, Bosque aleatorio y Optimización mínima secuencial.
- **Resultados**: El clasificador con mejores resultados en ambas bases fue la optimización mínima secuencial. En particular, para la base Emo_voz_mx1, usando la función ventana Blackman y una longitud de ventana de 25 milisegunos se obtuvo un porcentaje de 78.5 %. Para el caso de *EmoWisconsin*, usando la función Hamming y una ventana de 35 milisegundos se obtuvo un porcentaje de 44.9 %.

Además de los trabajos anteriores, en esta sección se reportarán tres trabajos de tesis de tres universidades internacionales encontrados en el motor de búsqueda de Google Académico usando la frase "*voice gender recognition thesis*"[1].

Primera Tesis: Latasha Roberts, "*Design and development of a gender and language recognition system*" (2008). *Doctoral dissertation, Tennessee State University* [16]. (Diseño y desarrollo de un sistema de reconocimiento de género e idioma"(2008). Tesis doctoral, Universidad Estatal de Tennessee").

- **Objetivo**: Se implementó un sistema de reconocimiento de género y lenguaje mediante características de la voz en ambientes controlados.

[1] Se traduce en Tesis de reconocimiento de género por voz.

- **Metodología**: El conjunto de datos consistió en 2300 muestras de audio provenientes de cinco sujetos. Cada sujeto hablaba un idioma diferente (inglés, español, alemán, farsi y arábigo). Además, cada uno de ellos grabó frases de su idioma en dos ambientes diferentes. Para la detección de género se extrajeron características cepstrales de los audios y se implementó una técnica de coincidencia de patrones utilizando la distancia euclidiana para reconocer el género. En el caso del algoritmo de detección de idioma, se implementó la distancia de Mahalanobis como la técnica de coincidencia de patrones para reconocer las características de una frase específica del idioma dado.
- **Resultados**: Véase la tabla 2.1.

Tabla 2.1: Resultados de la exactitud del reconocimiento de lenguaje de "Diseño y desarrollo de un sistema de reconocimiento de género e idioma". Fuente: [16].

	Inglés	**Español**	**Alemán**	**Farsi**	**Arábigo**
Tasa de reconocimiento - entrenamiento	79 %	65 %	82 %	71 %	77 %
Tasa de reconocimiento - prueba	87 %	75 %	88 %	80 %	83 %
Máximo valor en una frase	99 %	82 %	94 %	89 %	90 %

Segunda Tesis: Hassam Ullah Sheikh, *"Who is speaking? Male or female"* (2013). *Graduate Thesis and Dissertations. University of Manchester* [17]. ("¿Quién está hablando? Hombre o mujer"(2013). Tesis de grado y Disertaciones. Universidad de Manchester).

- **Objetivo**: Se implementó un sistema de reconocimiento de género por voz sin restricción de edad, lenguaje, acento ni dialecto en ambientes reales.
- **Metodología**: El esquema general de su modelo puede verse en la figura 2.1. Las ideas generales fueron las siguientes:
  - **Conjunto de datos**: 20 voces (10 mujeres y 10 hombres) con audios de una duración de 6 minutos cada uno provenientes del Corpus de voz de *Voxforge*. El lenguaje de cada voz corresponde a uno de los 12 idiomas estudiados (inglés, urdu, árabe, italiano, español, francés, alemán e idiomas regionales indios como bengalí, telugu, tamil, marathi y gujarati). Siendo el inglés el de mayor predominancia.
  - **Extracción de características**: Los audios fueron estudiados mediante la *toolbox* (término equivalente a librería en python) *Voice box*, se realizó la manipulación de los archivos de audio. En primer lugar, se utilizó las funciones *specsub* y *vodsohn* para eliminar la mayor cantidad de ruido de los audios. Seguidamente, se extrajeron las características de tono, coeficientes cepstrales de frecuencias de mel (*Mel Frequency Cepstral Coeficient o MFCC*) y coeficientes delta cepstrales desplazados (*Shifted Delta Cepstral o SDC*) en diferentes segmentos de audio (25 milisegundos para el tono y 30 milisegundos para ambos coeficientes).
  - **Clasificadores**: Se implementaron dos clasificadores. El primero fue una máquina de soporte vectorial usando las características del tono. Mientras que el segundo fue un modelo de mezclas gaussianas usando los coeficientes delta cepstrales desplazados.

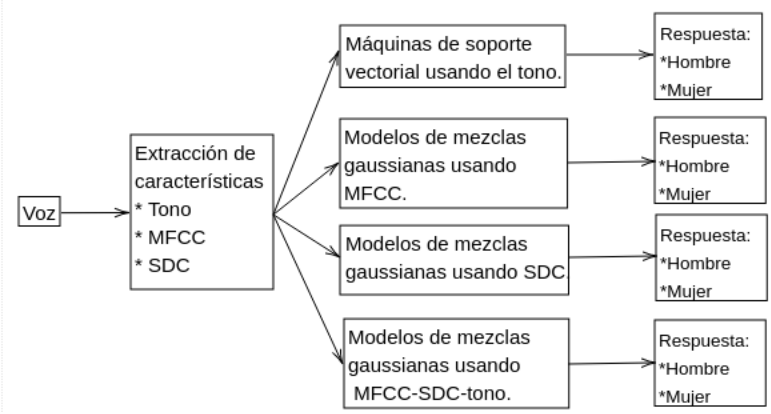


Figura 2.1: Esquema general de los 4 sistemas de reconocimiento de género por voz propuestos en [17].

- **Resultados**: Los resultados de la precisión de las simulaciones en los conjuntos de prueba se presentan en la tabla 2.2.

Tabla 2.2: Resultados de la precisión de los 4 sistemas de reconocimiento de género por voz propuestos con configuraciones diferentes presentados en [17].

Modelo	**Hombre**	**Mujer**
Máquinas de soporte vectorial con tono	73.95 %	85.93 %
GMM de 8 componentes con MFCC	57.37 %	72.02 %
GMM de 16 componentes con MFCC	59.31 %	74.14 %
GMM de 32 componentes con MFCC	65 %	74.64 %
GMM de 8 componentes con SDC	69.77 %	73.25 %
GMM de 16 componentes con SDC	72.03 %	74.60 %
GMM de 32 componentes con SDC	74.29 %	77.41 %
GMM de 8 componentes con MFCC, SDC y tono	86.80 %	70.10 %
GMM de 16 componentes con MFCC, SDC y tono	79.28 %	79.08 %
GMM de 32 componentes con MFCC, SDC y tono	79.28 %	80.33 %

Tercera Tesis: Erokyar Hasan, "*Age and Gender Recognition for Speech Applications based on Support Vector Machines*" (2014). *Graduate Thesss and Dissertations. University of South Florida*. [18]. (Reconocimiento de edad y género para aplicaciones de voz basadas en máquinas de vectores de soporte"(2014). Tesis de grado y Disertaciones. Universidad de Florida del Sur).

- Objetivo: Se implementó un sistema robusto de reconocimiento de edad y género para aplicaciones de voz que también provea buenas tasas de reconocimiento en condiciones del mundo real.
- Metodología: La propuesta general puede verse en la imagen 2.2. A grandes rasgos sus elementos son los siguientes.
  - **Conjunto de datos**: Se utilizó el conjunto de datos de voces en inglés para reconocimiento de voz (*English Language Speech Database for Speech Recognition o ELSDSR*). Conteniendo 22 hablantes, donde 10 de ellos son mujeres y el resto son hombres.
  - **Extracción de características**: Por cada audio se extrajeron dos características, el tono y los coeficientes cepstrales de frecuencias de Mel, en intervalos de 25 milisegundos.
  - **Clasificadores**: El clasificador utilizado fue máquina de soporte vectorial. Las etiquetas propuestas para el clasificador fueron: adulto joven (*Young Adult Male*), adulta joven (*Young Adult Female*), adulto mayor (*Middle Adult Male*) y adulta mayor (*Middle Adult female*). El primer grupo se delimitó de los 20 a los 40 años y el segundo grupo de los 40 a los 65 años.

- **Resultados:** Los resultados de su modelo principal pueden verse en la tabla 2.3.

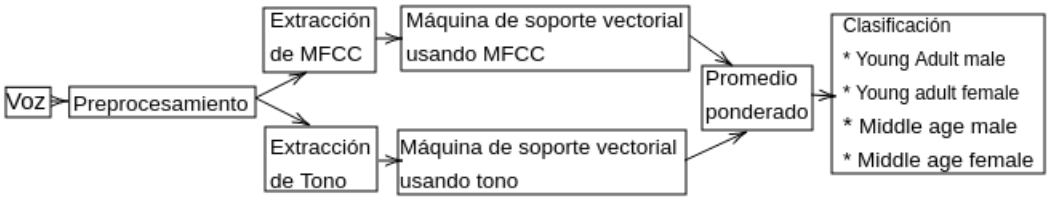


Figura 2.2: Esquema general del sistema de reconocimiento de edad y género mediante voz propuesto en [18].

Tabla 2.3: Resultados de la precisión del sistema de reconocimiento de género por voz presentados en [18].

	Precisión
Adulto joven	91.83 %
Adulta joven	67.67 %
Adulto mayor	24.17 %
Adulta mayor	73.15 %

2.2. Estado del Arte

El tema de reconocimiento de género mediante la voz ha sido estudiado ampliamente tanto en la perspectiva médica, matemática y computacional desde 1990 (véase [19–21]). Por lo anterior, para identificar ciertos artículos clave en su desarrollo y las actuales tendencias, se requirió de realizar una búsqueda elaborada de referencias. Delimitándolos mediante los siguientes núcleos temáticos del trabajo:

- Detección de características y sus técnicas de extracción mediante el análisis del audio.
- Elección de los clasificadores para determinar el género de una persona mediante las características elegidas.

2.2.1. Sintesis Tabular de las Referencias

Las tablas 2.4, 2.5, 2.6, 2.7 y 2.8 presentan los aspectos relevantes de los artículos leídos del 2018 al 2022. Las síntesis textuales de los trabajos se encuentran en el anexo: *Síntesis de artículos representativos del 2018 al 2022*. Cada tabla consiste en 5 columnas donde:

- La primera (Artículo) detalla el nombre de la referencia.
- La segunda (Objetivo) expone el objetivo de dicha investigación.
- La tercera (Características), emplea una palabra clave que detalla el tipo de características empleadas en el reconocimiento de género. Cepstrales, hacen referencia a características extraídas del espacio Cepstral. Mientras que acústicas detallan las características del espacio de Frecuencias.
- La cuarta (Conjuntos) muestra el conjunto de datos fue empleado.
- La quinta (Resultados) muestra las métricas obtenidas en dicha investigación.

Tabla 2.4: Artículos representativos del 2018.

2018				
Artículo	**Objetivo**	**Características**	**Conjuntos**	**Resultados**
On the Performance of Cepstral Features for Voice-Based Gender Recognition [22]	Analizar los desempeños de múltiples conjuntos de características cepstrales de tiempo corto con variación en el número de sus dimensiones en un sistema de reconocimiento de género.	Cepstrales	Primer conjunto de voces: *English language speech for speaker recognition (ELSDR)* de 197 audios. Segundo conjunto de voces: *Speaker in the Wild (SITW)* de 300 audios.	Los coeficientes cepstrales inversos brindaron mejor exactitud en comparación con las otras combinaciones.
Voice gender detection using gaussian mixture model [23]	Diseñar e implementar un sistema para la codificación, análisis, síntesis e identificación de género del hablante mediante coeficientes cepstrales y usando el clasificador de mezclas gausianas.	Cepstrales	Conjunto de voces Audio set corpus con 50 voces.	La exactitud del clasificador fue de 76 % para hombres. 95 % para mujeres.
A stacked technique for gender recognition though voice [24]	Diseñar e implementar un algoritmo de aprendizaje automático apilado para determinar el género utilizando los parámetros acústicos y construido a partir de los árboles de clasificación y regresión, máquinas de soporte vectorial y red neuronal.	Acústicos	El conjunto de voces proviene de *VoxForge* y *Festvox* teniendo 3160 muestras.	El método apilado obtuvo una exactitud del 97.05 %.
Comparison of Different Normalization Techniques on Speakers' Gender Detection [25]	Comparar los resultados obtenidos mediante normalizadores de vectores de características de frecuencias cepstrales en un clasificador de máquina de soporte vectorial para la distinción de género por voz.	Cepstrales	El conjunto de Voces es *TIMIT* con 192 muestras de audio (hombres y mujeres) Las normalizaciones fueron: * Normalización de media y varianza a corto plazo. * Tiempo cepstral medio y normalización de Escala. * Normalización Min-Max. *Desviación estándar.	Las normalizaciones tuvieron un efecto negativo en la exactitud para la categoría de mujer. Pero tuvieron un efecto positivo para la categoría de hombre.

Tabla 2.5: Artículos representativos del 2019.

2019				
Artículo	**Objetivo**	**Características**	**Conjuntos**	**Resultados**
An effective gender recognition approach using voice data via deeper LSTM networks [26]	Implementar un sistema computacional que realice predicciones del género de una persona mediante su voz, por medio de una red de memoria profunda de corto y largo plazo (*Deeper Long Short Term Memory*).	Acústicas	Conjunto de datos en Kaggle con 3168 archivos de audio.	Precisión: 98.4 % Sensibilidad: 97.2 % Especificidad: 99.5 % Media geométrica de la sensibilidad y la especificidad: 98.3 %
DGR: Gender Recognition of Human Speech Using One-dimensional Conventional Neural network [27]	Implementar una red neuronal que realice predicciones del género de una persona mediante 5 características cepstrales de la voz (espectrograma de Mel, Coeficientes cepstrales de frecuencias de Mel, transformada de Fourier de tiempo corto tipo cromagrama, Constraste espectral y tonnetzs).	Cepstrales	Conjunto de datos privado de archivo de audio de voces artificiales con 20 idiomas. Cada idioma consiste en 16 muestras de audio dividido en 8 voces masculinas y el resto femeninas.	Exhaustividad: 99.7 %
Voice based Gender Recognition [28]	Diseñar un sistema de reconocimiento de género mediante la implementación de 4 algoritmos de clasificación: Árboles de decisión, Potenciación del gradiente, Bosques aleatorios y máquinas de vectores de soporte. Eligiendo él de mejor desempeño por medio de la métrica de precisión.	Acústicas	El conjunto de datos recopilados de *VoxForge* consta de 62440 muestras de audio comprimidas y divididas en conjuntos de 10 archivos.	Precisión de los clasificadores implementados. Árbol de decisión: 86.9 % Potenciación del gradiente: 93.7 % Bosques aleatorios: 89.3 % Máquinas de soporte vectorial: 90.5 %
Gender Identification Via Voice Analysis [29]	Diseñar un sistema de reconocimiento de género mediante la agrupación e implementación de 4 modelos de clasificación CART, XGBoost, SVM y *Random Forest* por medio de parámetros acústicos de voz.	Acústicos	Un conjunto de datos privado de 3000 muestras de voz.	Precisión de los clasificadores implementados: Regresión logística clásica: 50 % Análisis de regresión logística completo: 71 % Árbol de clasificación y regresión: 79 % Máquinas de vectores de soporte: 86 % Bosque aleatorio: 88 % Potenciación del gradiente: 88 % Modelo ensamblado: 89 %

Tabla 2.6: Artículos representativos del 2020.

2020				
Artículo	**Objetivo**	**Características**	**Conjuntos**	**Resultados**
Gender Detection From Human Voice Tensor Analysis [30]	Diseñar e implementar un sistema de reconocimiento de género mediante el análisis de tensores.	Cepstrales	*TIMIT-DR1*, *TIMIT-Mix* y *SHRUTI*.	A mayor cantidad de entradas en el vector cepstral y de eigen vectores se obtuvo una mejor precisión en el sistema.
Gender determination Using Voice Data [31]	Diseñar e implementar una red neuronal de tipo perceptrones multicapa para el reconocimiento de género mediante parámetros acústicos.	Acústicos	Kaggle	Métricas del clasificador: Exactitud: 97.9 % Exhaustividad: 98 % Precisión: 97.7 % Puntaje F1: 97.9 %
Voice Gender Recognizer- Recognition of Gender from Voice using DNN [32]	Construir un modelo para la predicción de género mediante audios de voz basado en redes neuronales profundas.	Acústicas	Kaggle	El modelo de perceptrones multicapa obtuvo mejores resultados, teniendo 96 % de exactitud. No se tuvieron problemas de sobre ajuste ya que se usaron las técnicas de "*dropout*" y "*batch normalization*".
Gender identification from arabic speech using ML [33]	Diseñar y desarrollar un sistema de reconocimiento de género y estimación de edad basado en la extracción de características MFCC del habla árabe. Utilizando seis algoritmos de aprendizaje como árbol de decisiones, bosque aleatorio, K-vecinos más cercanos, máquinas de soporte vectorial, redes neuronales artificiales y bayes ingenuo.	Cepstrales	Corpus *Urban Jordan* (6 hombres y mujeres).	La máquina de soporte vectorial y la red neuronal de tipo perceptrones multicapa fueron superiores en la determinación de género con precisiones del 98.5 % y 96.5 % respectivamente. Los árboles de decisión y bosque aleatorio fueron superiores en la estimación de la edad con precisiones del 95.9 % y 93.0 %.

Tabla 2.7: Artículos representativos del 2021.

2021				
Artículo	**Objetivo**	**Características**	**Conjuntos**	**Resultados**
Deep learning of Voice Gender Idenfication [34]	Diseñar una red convolucional para clasificar el género de una voz mediante el espectrograma de Mel obtenido de esta.	Cepstrales	Conjunto de datos privados de 278 voces de hablantes masculinos y femeninos.	Precisión: 93 % Exhaustividad: 95 % Puntaje F1: 94 %
Performance Analysis of ML Algorithms to Detect Gender Based on Voice [35]	Analizar el rendimiento de 5 algoritmos clasificadores de género por voz: bosque aleatorio, árbol de decisión, máquina de soporte vectorial, red neuronal y potenciación del gradiente.	Acústicas	Conjunto de 3000 Audios de voz provenientes de *VoxForge*.	Mejor exactitud: Potenciación del gradiente Mejor precisión: Red neuronal
A comparative Study of Deep learning and Machine Learning Approaches in Speech Emotion and Gender Recognition System [36]	Examinar el desempeño del aprendizaje por máquina y el aprendizaje profundo en la detección de género y emociones (feliz, enojado, neutral y triste).	Cepstrales	Dos conjuntos de datos construidos: Voz inglesa: 192 hombres y 214 mujeres. Voz canarés: 136 hombres y 146 mujeres.	Métricas solo de género: Inglés: Exactitud: 97.1 % Precisión: 97.9 % Exhaustividad: 95.8 % Canarés: Exactitud: 84 % Precisión: 85.4 % Exhaustividad: 82 %
Voice gender recognition under unconstrained environments using self-attention [37]	Diseñar e implementar dos modelos de reconocimiento de género basado en atención propia.	Cepstrales	Se construyeron dos subconjuntos a partir del conjunto de voces de *VoxCeleb* Uno de 12000 audios y el segundo de 25000.	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 2.8: Artículos representativos del 2022.

2022				
Artículo	**Objetivo**	**Características**	**Conjuntos**	**Resultados**
Age group classification and gender recognition from speech with temporal cnn [38]	Evaluar el desempeño de 18 redes neuronales de los tipos convolucional y convolucional temporal de diversas arquitecturas, para determinar una configuración óptima en el problema de detección de género y rango de edad en sistemas de respuesta de voz interactivos	Cepstrales	"*Mozilla Common Voice*" en el idioma inglés.	Los resultados confirman que todos los tipos de redes obtienen resultados significativos para la clasificación de género, pero la combinación de convolucional y convolucional temporal obtuvieron mejores métricas.
NeuraGen-A Low-Resource Neural Network based approach for Gender Classification [39]	Diseñar e implementar una red neuronal de tipo perceptrones multicapa para el reconocimiento de género mediante parámetros acústicos	Cepstrales	*ELSDSR* y *TIMIT*.	La red neuronal propuesta obtuvo resultados significativos en cada una de sus métricas (mayores a 50 %). Por lo que es una primera versión para el diseño de redes neuronales de bajos recursos.
Gender Identification in a Two-Level Hierarchical Speech Emotion Recognition System for an Italian Social Robot [40]	Diseñar e implementar una red neuronal de bajos recursos que permita realizar la automatización de la detección de género por voz.	Cepstrales	Se utilizó la base de datos emocional italiana *EMOVO*.	El sistema propuesto escucha continuamente el entorno proporcionando la información de género del hablante con alta precisión.
Speaker Gender Recognition Based on Deep Neural Networks and ResNet50 [41]	Desarrollar un sistema *Speech Emotion Recognition (SER)* en robots sociales para el monitoreo de pacientes hospitalizados y residentes en el hogar.	Cepstrales	Se utilizó una sub-base de 6995 audios de hombres y 5662 audios de mujeres del sitio web *Common Voice*.	Las redes neuronales propuestas tuvieron métricas superiores al 90 %.

2.2.2. Análisis de los Artículos Sintetizados

Una manera de entender el desarrollo de los sistemas de reconocimiento de género mediante análisis acústico de voz, es a partir de los siguientes puntos:

- Características que utilizan en el clasificador.
- Lenguajes de la voces.
- Ambiente de grabación de la voz y multietiquetas.

Características que se utilizan en el clasificador

Las características empleadas en los trabajos resumidos son acústicas o cepstrales. Las características acústicas se analizan con técnicas estadísticas clásicas. Por ejemplo, se utilizan diagramas de dispersión, distribuciones de probabilidad y matrices de correlación para cada género (véase los artículos [26], [29] y [32]). Además, a cada parámetro acústico se le brinda un peso w obtenido por los clasificadores implementados (véase [26] y [28]). Por otro lado, los artículos enfocados en características acústicas utilizan múltiples clasificadores y técnicas de ensamblado para resolver el problema de la clasificación (véase [31] y [35]). De este modo se evitan problemas de sobreajuste.

Los trabajos que emplean características cepstrales usualmente emplean algoritmos de aprendizaje profundo. En [22], se exploraron múltiples combinaciones en una red neuronal de perceptrones multicapa. Mientras que en [25], se compararon múltiples técnicas de normalización para los coeficientes cepstrales para mejorar las métricas del clasificador. Se han considerado más características espectrales (véase [27]) para la detección de género. Además, por su forma vectorial, se han implementado con análisis tensorial (véase [30]). Sin embargo, las tendencias de los últimos años han sido diseñar redes más sofisticadas (por ejemplo, temporalmente convolucionales) y utilizando varios miles de coeficientes para mejorar la precisión del reconocimiento de género (véase alguno de los ejemplos de la tabla 2.8) y detección de emociones (véase [36]).

Lenguajes de la voces

El lenguaje más estudiado en el reconocimiento de género en los artículos reportados fue el inglés. Sin embargo, algunos de estos también trabajan con un segundo lenguaje (véase [27], [33] y [36]).

Ambiente de grabación de la voz y multietiquetas

Las voces que utiliza un sistema de reconocimiento de género pueden estar en ambientes controlados y no controlados. En el primer caso, se utilizan micrófonos especializados, cada audio dura lo mismo y cada voz reproduce un texto específico (véase [34] y [27]). En el segundo caso, las voces provienen de conjuntos de voces públicas, donde cada usuario puede subir su propio audio grabado. Un ejemplo de estas bases de datos son *Mozilla Common Voice Dataset* y *Vox Forge*. Además, debido al creciente uso del aprendizaje profundo, es posible extraer tanto el género como un rango de edad de la persona que habla (véase [38]).

3 | Marco Teórico

El presente capítulo presenta las bases teóricas de las tres disciplinas fundamentales en este trabajo Procesamiento de señales, modelación biomatemática y teoría de aprendizaje de máquina.

3.1. Marco Teórico Biológico y Matemático

La vibración de las cuerdas vocales es la fuente habitual de sonido en las vocales, y el tracto vocal es un filtro acústico que modifica el sonido producido por las cuerdas vocales (véase [42–44]). El número de veces que esta onda periódica compleja se repite por segundo, determina su frecuencia fundamental (F_0) y está relacionada con la percepción del tono (*pitch*) de la voz por parte del oyente. Los múltiplos de F_0 se denominan armónicos y son producidos mediante las diversas configuraciones que puede tener el tracto vocal. A esta explicación sobre la acústica del habla se conoce como la teoría de la fuente-filtro (véase [42, 45], las Figuras 3.1 y 3.2), y se resume en los siguientes puntos:

1. Se modela el tracto vocal como un tubo lleno de aire que está abierto en un extremo (la boca está abierta en su mayor parte durante el habla) y cerrado en el otro extremo (laringe).
2. La vibración de las cuerdas vocales es una fuente de energía que viaja como onda por el tracto vocal.
3. Se produce la resonancia del tracto vocal. Esto es, la vibración producida por las cuerdas es igual o cercana a la del tracto vocal.
4. Se generan las frecuencias formantes, las cuales son bandas de frecuencia donde se concentra la mayor parte de la energía sonora de la onda.
5. Estas resonancias o formantes son descritos según tres parámetros: el centro de frecuencia, ancho de banda y energía. Al modificar la forma del tracto vocal se modifican estos tres elementos en diferente medida. Como resultado, el sonido generado cambiará según estas modificaciones.
6. La menor de las frecuencias de los formantes tiene una longitud de onda (la distancia que recorre una perturbación periódica que se propaga por un medio en un ciclo) λ igual a cuatro veces la longitud del tracto vocal.
7. Los armónicos (múltiplos de F_0) provenientes del sonido laríngeo serán reforzados o atenuados por estas resonancias o formantes. De esta forma, los armónicos cercanos a los valores de las frecuencias formantes, serán más amplificados que los armónicos que se encuentren más alejados de los formantes.

A partir de los puntos anteriores, y de las siguientes hipótesis se puede modelar el tracto vocal (véase [46, 47]) y la producción del habla.

Tracto vocal
Conjunto de hipótesis:

a La forma del tracto vocal no cambia con el tiempo.
b No hay fricción de la onda con las paredes del tracto vocal.
c No hay pérdidas de energía debido a la viscosidad ni a las condiciones térmicas.
d Se considera que los tubos tienen un tamaño tal que sólo puede haber una frecuencia posible por tubo.

Conjunto de variables y parámetros:

- Variables:
  - Tiempo t medido en segundos.
  - Posición x en el eje del cilindro medido en centímetros.
  - Presión del sonido $p(x,t)$ dentro del tubo en el punto (x,t) medida en Pascales.
  - Velocidad del volumen del aire ($u(x,t)$) dentro del tubo en el punto (x,t), medida en centímetros

cúbicos por segundo.

- Parámetros:
  - Densidad del aire en el tubo, denotada por ρ.
  - Velocidad del sonido c en centímetros por segundo.
  - Área transversal del tubo A en centímetros cuadrados.

Modelación del tracto vocal (véase [46, 47]):

$$\left.\begin{aligned} -\frac{\partial p(x,t)}{\partial x} &= \frac{\rho}{A}\frac{\partial u(x,t)}{\partial t} \\ -\frac{\partial u(x,t)}{\partial x} &= \frac{A}{\rho c^2}\frac{\partial p(x,t)}{\partial t} \end{aligned}\right\} \quad \textbf{Modelo clásico del tracto vocal} \tag{3.1}$$

A partir de Ecuación 3.1 y las hipótesis establecidas, se infiere la siguiente ecuación (véase [42, 45, 47])

$$F_n = \frac{c}{\frac{4}{2n-1}L}, \tag{3.2}$$

donde n y F_n son el índice y la $n-$ésima frecuencia formante, respectivamente. La constante c es la velocidad del sonido dentro de la garganta (para este trabajo se asume $c = 35000\frac{cm}{s}$) y L es la longitud del tracto vocal. De este modo, despejando L de Ecuación 3.2, se tiene:

$$\left.L = \frac{c}{\frac{4}{2n-1}F_n}\right\} \quad \textbf{Estimación de la longitud del tracto vocal} \tag{3.3}$$

Producción del habla mediante la teoría fuente filtro:
Conjunto de hipótesis

- El sonido producido $s(t)$ por el habla al tiempo t se puede descomponer en dos componentes: la fuente y el filtro.
- La fuente $g(t)$ o señal glótica, es la parte responsable de la generación de sonidos y está formada por las cuerdas vocales y las cavidades supraglóticas.
- El filtro $v(t)$ es la parte responsable de dar forma y modificar los sonidos generados por la fuente. Está formado por el resto del tracto vocal.
- Para las señales en donde la vibración y posición de los labios modifican la señal acústica, se añade un segundo filtro denominado radiación labial $l(t)$.
- La convolución de $g(t), v(t)$ y $l(t)$ genera la señal $s(t)$.

Conjunto de variables

- $s(t)$ es la señal analógica del habla al tiempo t.
- $S(\omega)$ es el espectro de la señal del habla $s(\omega)$ en la frequencia ω.
- $l(t)$ es la radiación labial al tiempo t.
- $L(\omega)$ es el espectro de la radiación labial $l(t)$ en la frequencia ω.
- $v(t)$ es el filtro del tracto vocal al tiempo t.
- $V(\omega)$ es la respuesta frequencial del tracto vocal $v(t)$ en la frequencia ω.
- $g(t)$ es la señal glótica al tiempo t.
- $G(\omega)$ es el espectro de la fuente glótica $g(t)$ en la frequencia ω.
- La frecuencia fundamental (F_0): es la frecuencia más baja en la que vibran las cuerdas vocales (en Hertz).
- Periodo de las cuerdas vocales (T): es el tiempo que tarda una vibración completa de las cuerdas vocales (en segundos).
- La longitud del tracto vocal (L): es la distancia desde las cuerdas vocales hasta la abertura de la boca (en centímetros).

Modelación de la producción del habla mediante la teoría fuente filtro (véase la Fig. 3.2):

$$s(t) = (g(t) * v(t)) * l(t) \quad \Big\} \quad \textbf{Modelo fuente filtro} \tag{3.4}$$

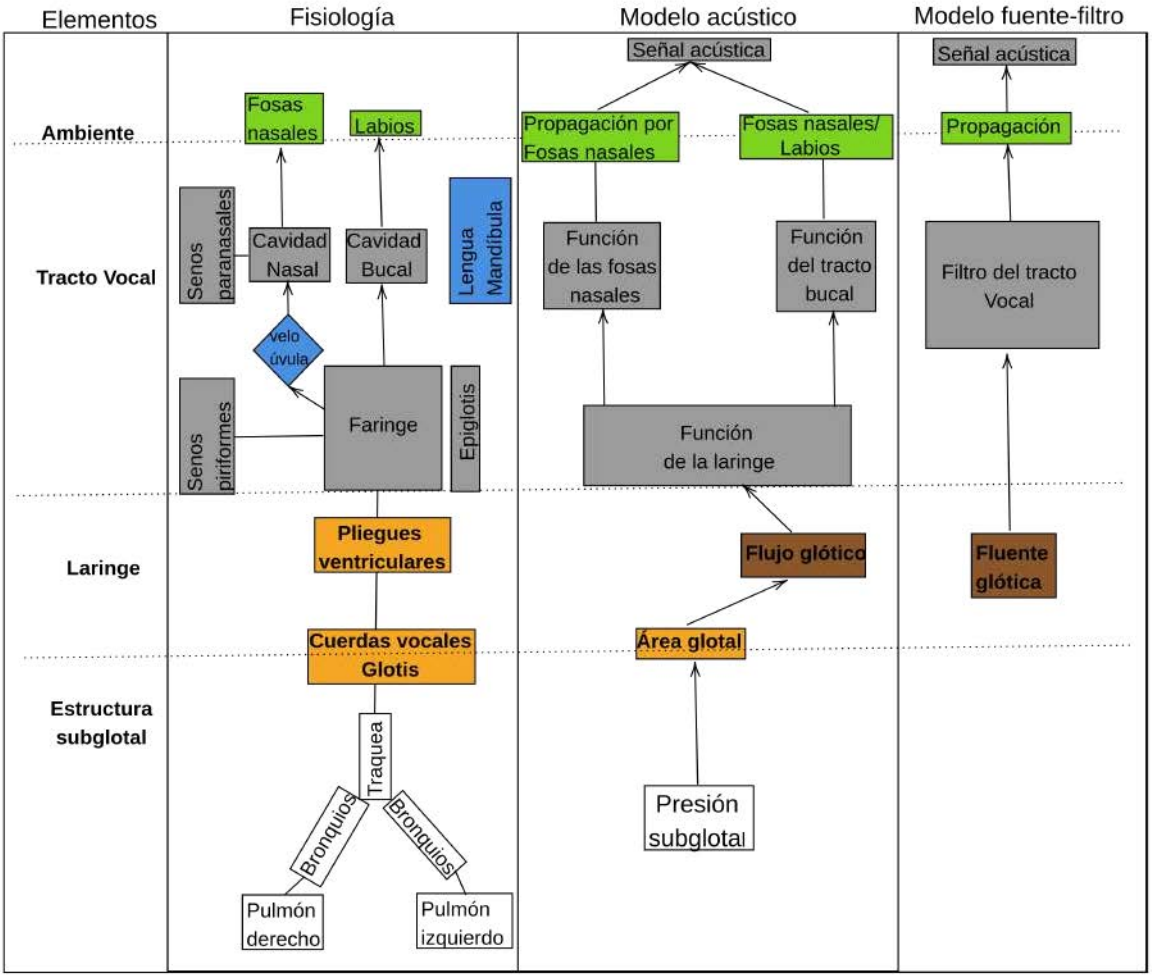


Figura 3.1: Modelación y esquema general de los elementos fisiológicos y acústicos del sistema fonatorio. Fuente: [63]

.

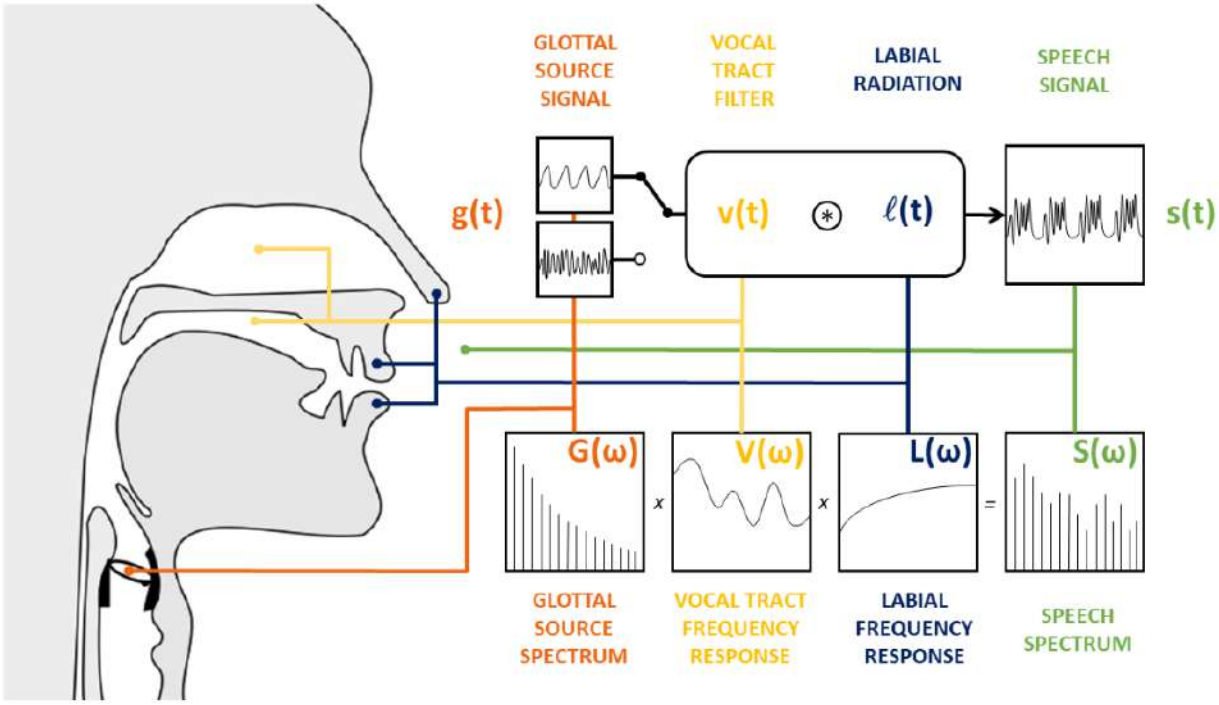


Figura 3.2: Esquematización del proceso de producción del habla basado en el modelo fuente-filtro y las diferentes contribuciones del sistema fonatorio al desarrollo de la señal del habla. Traducción: *Glottal source signal* (señal de fuente glotal), *Vocal tract filter* (filtro de tracto vocal), *Labial radiation* (radiación labial), *Speech signal* (señal del habla), *Glotal source spectrum* (Espectro del flujo glotal), *Vocal tract frequency response* (Respuesta frequencial del tracto vocal), y *Labial frequency response* (Respuesta frecuencial labial). Fuente: [48]

3.2. Marco Teórico Computacional

En esta sección se brindarán los conceptos relacionados al preprocesamiento de los audios. Así como, la extracción de las características acústicas y cepstrales de manera general y no de una metodología en específico. Se ha decidido exponerlo de esta manera para que se pueda explicitar el formalismo matemático subyacente de cada técnica y concepto. Se inicia con dos filtros de preprocesamiento (substracción espectral y de actividad de voz). Se prosigue con la extracción de características acústicas, cepstrales y finalmente de intensidad de voz.

3.2.1. Filtros Relacionados con Preprocesamiento de Voz

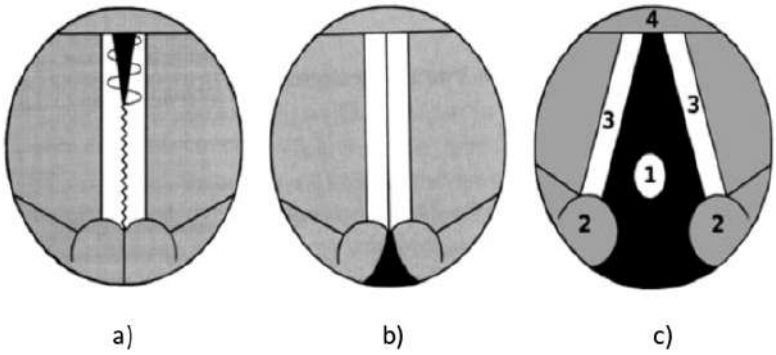
1. **Substracción espectral**
   La atenuación de ruido mediante substracción espectral es una técnica comúnmente utilizada para reducir el ruido de fondo en señales de audio (véase [49]). El principio básico de esta técnica es que la señal de audio que contiene el ruido de fondo se puede descomponer en diferentes bandas de frecuencia utilizando una transformada de Fourier. Luego, se puede estimar el espectro de potencia del ruido de fondo en cada banda de frecuencia utilizando técnicas de estimación de ruido. Este espectro de potencia estimado se resta del espectro de potencia de la señal original en cada banda de frecuencia para obtener una señal de audio limpia con menos ruido de fondo. El filtro *Remove Noise Spectral Subtraction* en PRAAT[1] (véase [49]), es un filtro de preprocesamiento utilizado en el análisis de señales de audio para eliminar el ruido de la señal. La formulación matemática del filtro *Spectral Subtraction* es la siguiente:

$$S'(f) = \text{máx}(S(f) - N(f), 0), \tag{3.5}$$

   donde $S(f)$ es la energía espectral de la señal original en la frecuencia f, y $N(f)$ es la energía espectral del ruido estimado en la frecuencia f. $S'(f)$ es la energía espectral de la señal filtrada en la frecuencia f. El filtro funciona restando la energía espectral del ruido estimado de la energía espectral de la señal original y tomando el máximo entre este valor y cero para evitar valores negativos. Los parámetros del filtro *Spectral Subtraction* en PRAAT incluyen:
   - Tamaño de la ventana de análisis: es el tamaño de la ventana utilizada para calcular la energía espectral de la señal y el ruido estimado. Este parámetro se puede ajustar según la longitud de la señal y la frecuencia de muestreo para obtener una estimación precisa del ruido.
   - Método de estimación de ruido: PRAAT ofrece varios métodos para estimar el ruido de la señal, por ejemplo el cálculo de la energía media, la estimación del ruido mediante una muestra de la señal o la estimación de ruido mediante una sección silenciosa.
   - Relación señal-ruido (SNR): Este parámetro establece el nivel de filtrado que se aplicará a la señal. Un valor más alto de SNR producirá una mayor eliminación de ruido, pero también puede eliminar información de la señal.

[1] PRAAT es un software especializado en el análisis acústico diseñado por el Dr. Paul Boersma. Para más información véase [53].

2. **Detector de actividad de voz**.
   Posteriormente a este proceso, se procede con la descomposición de la señal en fragmentos sonoros y no sonoros (en Inglés *Extract Voiced/Unvoiced Speech* del *Vocal Toolkit* [70] en PRAAT). La extracción de sonidos sonoros (*voiced speech*) es una técnica utilizada para separar los sonidos producidos por la vibración de las cuerdas vocales de los sonidos no vocales en señales de audio [69]. Los sonidos sonoros se caracterizan por tener una estructura rítmica y repetitiva, lo que los hace distinguibles de otros sonidos (véase la Fig. 3.3). Esta técnica se basa en el análisis del espectro de la señal de audio y la identificación de las regiones del espectro que corresponden a los sonidos sonoros. En general, se utiliza un algoritmo de análisis de Fourier para obtener la frecuencia fundamental de la señal de audio. A continuación, se utiliza un filtro digital para separar los sonidos sonoros de los sonidos no vocales. Los sonidos no vocales, como el ruido de fondo y los sonidos sibilantes, se pueden eliminar mediante técnicas de filtrado adicionales.



(i) Ilustración de las diferentes configuraciones de las cuerdas vocales que dan lugar a distintas formas de fonación: a) voz normal (*voiced speech*; b) voz susurrada; c) sin voz (*unvoiced speech*). Las estructuras numeradas corresponden a: 1. glotis; 2. Cartílagos aritenoides; 3. cuerdas vocales; y 4. epiglotis. Fuente: [48].

(ii) Secuencia de fotogramas de videoendoscopia de alta velocidad. Fuente: [48].

Figura 3.3: Representación del movimiento de cuerdas vocales (Figura superior (i)) e imágenes de una videoendoscopia (Figura inferior (ii)).

La formulación matemática del algoritmo se basa en la estimación del espectro de potencia de la señal de voz y la aplicación de un umbral para determinar las regiones de habla sonora (voz o *voiced speech*) y no sonora (silencio o *unvoiced speech*). La señal de voz se considera sonora si el espectro de potencia está por encima de un umbral determinado, y no sonora si está por debajo de ese umbral.

3.2.2. Extracción de Características Estadísticas del Tono

La señal digital de la voz y se compone esencialmente de dos elementos: parte sonora (*voiced speech*) y parte no sonora (*unvoiced speech*) [59]. El primer componente se asume como una función pseudoperiódica[2] x

[2]Véase [60] para la definición de pseudoperiódica

y el segundo componente como un ruido estacionario[3] r. Formalmente, se tiene

$$\forall n \in \mathbb{Z} : s(n) = x(n) + r(n). \tag{3.6}$$

En términos prácticos, se denomina a la parte cuasiperiódica como sonora (hay presencia de vibración de cuerdas vocales) y a la parte de ruido como no sonora (no hay vibración de las cuerdas vocales). Si bien, esta hipótesis es válida debido a la biología propia de las cuerdas vocales y del tracto vocal, la estimación del periodo de x es un problema complejo. Para este trabajo, se estima el periodo de x mediante la función de autocorrelación con retraso τ (véase [62]).

Función de autocorrelación con retraso τ: Sea s, una señal digital de periodo de muestreo T y dominio el intervalo [0,T]. La función de autocorrelación $r_s(\tau)$ con retraso $\tau > 0$ se define como:

$$r_s(\tau) = \sum_{t=0}^{T-1} s(t)s(t+\tau). \tag{3.7}$$

La función r_s mide la similitud entre una señal s y la misma señal desplazada por τ. A partir de esta observación se tiene que

- La función r_s tiene un máximo global cuando $\tau = 0$.[4]
- Si existe un máximo τ de r_s no nulo, entonces se dice que s es periódica y de periodo τ. De este modo, la frecuencia fundamental F_0 de s se define:

$$F_0 = \frac{1}{\tau}. \tag{3.8}$$

- Si r_s si tiene máximos locales en diversos $\tau_1, ..., \tau_n$, entonces se dice que la señal tiene una parte periódica o que es pseudoperiódica. Los recíprocos de esos τ_i son la frecuencia fundamental en ese subintervalo.

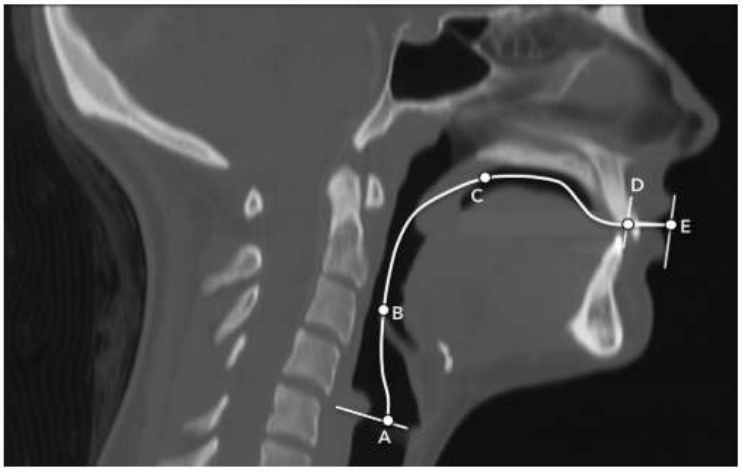
El algoritmo que utiliza PRAAT para calcular la frecuencia fundamental (también conocida como frecuencia fundamental de la voz, F_0, *pitch* o tono) se basa en el método de análisis de autocorrelación (véase [62]). Los pasos son los siguientes:

1. Se divide la señal vocal en pequeñas ventanas de tiempo (típicamente de 20-30 ms de duración) y se aplica una ventana de Hamming para reducir el efecto de las discontinuidades en los bordes de la ventana.
2. Se calcula la función de autocorrelación de cada ventana. La función de autocorrelación mide la similitud entre la señal en un momento dado y la señal retardada un cierto número de muestras de tiempo. La autocorrelación es máxima para la frecuencia fundamental y sus múltiplos enteros.
3. Se aplica un filtro para eliminar las altas frecuencias que no están relacionadas con la frecuencia fundamental de la voz.
4. Se busca el primer máximo de la función de autocorrelación dentro de un rango de frecuencias de interés (generalmente entre 75 y 600 Hz) y se toma el período correspondiente a ese máximo como la frecuencia fundamental.

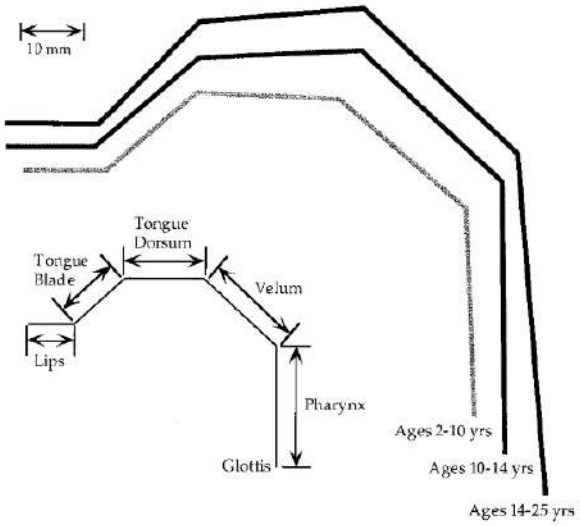
3.2.3. Extracción de los Estadísticos de los Primeros Cuatro Formantes y Estimación de la Longitud del Tracto Vocal

La extracción de la longitud del tracto vocal es una técnica utilizada para estimar la longitud del tracto vocal humano a partir de señales de audio (véase [67]). La longitud del tracto vocal se refiere a la distancia entre las cuerdas vocales y la boca teniendo un efecto significativo en la producción de sonidos vocales (véase Fig. 3.4).

[3] Véase [61] para la definición de ruido estacionario.
[4] La demostración se sigue de la desigualdad de Cauchy.



(i) Ilustración de los puntos de referencia anatómicos medidos a lo largo del perfil de la línea central del tracto vocal (A glotis, B cara superior de la epiglotis, C unión de los paladares, D cara posterior de los labios superior e inferior a nivel del estomión, y E cara anterior de los labios superior e inferior a nivel del estomión. Fuente: [68].



(ii) Morfología media del tracto vocal (medida con el método del "segmento lineal") en niños, adolescentes y adultos. Traducción: *Lips* (labios), *Tongue Blade* (punta de la lengua), *Tongue Dorsum* (dorso de la lengua), *Velum* (velo), *Pharynx* (faringe), *Glottis* (glotis). Fuente: [64].

Figura 3.4: Representación del tracto vocal mediante resonancia magnética y estimación lineal a escala.

La técnica de extracción de la longitud del tracto vocal se basa en el análisis del espectro de la señal de audio y la identificación de las frecuencias de resonancia del tracto vocal. Estas frecuencias de resonancia se denominan formantes, y son determinadas por la forma y longitud del tracto vocal. La expresión que se utilizó para su estimación puede verse en la ecuación 3.3.
El algoritmo que utiliza PRAAT para extraer los formantes de la voz se basa en el análisis de la señal de sonido mediante el uso de Transformada de Fourier de tiempo corto (*Short Time Fourier Transform* o STFT por sus siglas en inglés. Para más información véase [47, 53, 62, 69]). A partir de la *STFT*, PRAAT utiliza un algoritmo de análisis de formantes basado en un modelo de resonador de tubo con un filtro de formantes de orden n (véase [47]). Para esta investigación se utilizó $n = 4$. La extracción de los formantes de la voz en PRAAT se lleva a cabo a través de los siguientes pasos:

1. Preénfasis de la señal: se realiza una preénfasis de la señal para amplificar los formantes y reducir la influencia de los armónicos.
2. Análisis de la señal: se divide la señal de sonido en pequeñas ventanas de tiempo (*frames*) de duración fija y se aplica la STFT a cada ventana para calcular la frecuencia y la amplitud de los componentes espectrales.
3. Selección de los picos espectrales: se seleccionan los picos espectrales (máximos locales) dentro de cada ventana (*frame*) de la señal.
4. Estimación de los formantes: se utiliza un algoritmo de búsqueda de formantes que se basa en un modelo de resonador de tubo para estimar los formantes de la señal. Este algoritmo busca la combinación de formantes que mejor se ajusta a los picos espectrales encontrados en el paso anterior.
5. Filtrado de formantes: se aplica un filtro de formantes de orden n para suavizar las estimaciones de los formantes y eliminar los errores de medición.
6. Refinamiento de los formantes: se realizan ajustes finos a los formantes para mejorar su precisión y eliminar los errores de medición.

El resultado de este proceso es un conjunto de formantes estimados para cada *frame* (ventana) de la señal de sonido, lo que permite analizar la variación de los formantes a lo largo del tiempo y extraer información sobre las características de la voz, como la altura y la forma de la vocal.

3.2.4. Extracción de Estadísticos de la Intensidad de la Voz

La extracción de la intensidad de la voz vocal es una técnica utilizada para medir la energía acústica de la señal de audio producida por las cuerdas vocales (véase [66]). La intensidad de la voz se refiere a la cantidad de energía acústica que se transmite a través del aire durante la producción de sonidos vocales. La técnica de extracción de la intensidad de la voz se basa en el análisis del nivel de presión sonora de la señal de audio en diferentes frecuencias utilizando técnicas de análisis espectral, como la transformada de Fourier. La intensidad de la voz puede ser medida en términos de decibelios (dB), que es una unidad de medida para la energía acústica.

PRAAT utiliza el modelo matemático de energía acústica para calcular la intensidad de la voz (véase [53]). Este modelo se basa en el cálculo de la energía contenida en una señal de voz y se expresa en términos de decibelios (dB) de presión sonora. El cálculo de la intensidad en PRAAT se realiza mediante el siguiente procedimiento:

- Se divide la señal de voz en pequeños segmentos de tiempo (*frames*) de duración 25 milisegundos.
- Para cada segmento de tiempo, se calcula la energía acústica como la suma de los cuadrados de las amplitudes de la señal de voz en ese segmento.
- A partir de la energía acústica calculada en el paso anterior, se obtiene la intensidad en decibelios

utilizando la siguiente fórmula (véase [58]):

$$\text{Intensidad (dB)} = 10 \log_{10} \left(\frac{\text{Energía}}{\text{Área}} \right), \tag{3.9}$$

donde Energía es la energía acústica calculada en el paso anterior y Área es el área de la ventana de análisis utilizada para el cálculo de la energía. En este caso, el área a la que se refiere la fórmula de intensidad acústica no se refiere a un área física en el espacio, sino que se trata de un término abstracto que se utiliza para hacer referencia a la superficie sobre la cual se distribuye la energía acústica de la señal en la ventana de tiempo considerada. Por lo tanto, la intensidad de la voz se mide en función de la cantidad de energía acústica que se distribuye en esa ventana específica de la señal de voz. Es importante destacar que este modelo de intensidad se basa en la energía de la señal y no en la presión sonora directa.

3.2.5. Extracción de la Mediana de los Coeficientes Cepstrales de Frecuencias de Mel

La extracción de coeficientes cepstrales de frecuencias de mel en audios es una técnica ampliamente utilizada en el procesamiento de voz y audio (véase [65]). Esta técnica implica el cálculo de los coeficientes de la transformada de cepstrum, una representación en el dominio de la frecuencia de la envolvente espectral de una señal. En particular, los coeficientes cepstrales de frecuencias de mel se calculan a partir de la transformada de cepstrum de una señal de audio que ha sido previamente transformada en un espectrograma de frecuencias de mel. El espectrograma de frecuencias de mel es una representación logarítmica de la energía de la señal en diferentes bandas de frecuencia, que se ajustan de manera no uniforme a la escala de frecuencias perceptualmente relevantes para el oído humano. Los coeficientes cepstrales de frecuencias de mel pueden utilizarse para caracterizar las propiedades acústicas de una señal de audio, como su tonalidad, timbre y entonación, y se utilizan en una amplia gama de aplicaciones, como el reconocimiento automático de voz y el análisis de la música.

El proceso utilizado por PRAAT para extraer los coeficientes cepstrales de frecuencias de mel consta de los siguientes pasos (véase [53]):

Análisis de la señal de sonido: Se divide la señal de sonido en pequeñas ventanas de tiempo, de duración de 25 milisegundos y se les aplica una función de ventana para minimizar las discontinuidades en los bordes.

Transformada de Fourier: Se aplica una transformada de Fourier Rápida (*Short Time Fourier Transform* o STFT, por su nombre en inglés) a cada ventana de la señal. El resultado es una representación en el dominio de la frecuencia de la señal de sonido para cada ventana. La implementación de la transformada rápida de Fourier en PRAAT tiene la siguiente expresión

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{-j2\pi kn}{N}}, \ k \in \{0, ..., N-1\}. \tag{3.10}$$

Filtro en la escala de Mel: Se aplica un banco de filtros en la escala de Mel a cada resultado de la transformada de Fourier para enfatizar las bandas de frecuencia relevantes para la percepción auditiva humana (véase [12]). La escala de Mel se define como

$$\text{mel} = 2595 * \log_{10} \left(\frac{1+f}{100} \right), \tag{3.11}$$

donde f es la frecuencia de la señal en Hertz, y *mel* son los valores de la frequencia en la escala de Mel.

Transformada de Coseno: Se aplica una transformada de coseno inversa (también conocida como transformada de cepstrum) a cada resultado filtrado. La transformada de coseno se utiliza porque es más eficiente que la transformada de Fourier para procesar señales periódicas como la voz. El resultado de esta transformada es una representación en el dominio cepstral. Para esto, se define un banco de filtros (véase [57]), con $M \in \mathbb{N}$ elementos de la forma:

$$\forall m \in \{1, ..., M\} : H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{2(k-f[m-1])}{(f[m+1]-f[m-1])(f[m]-f[m-1])} & f[m-1] \leq k \leq f[m] \\ \frac{2(f[m+1]-k)}{(f[m+1]-f[m-1])(f[m+1]-f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & k > f[m+1] \end{cases} \tag{3.12}$$

La figura 3.5 es el gráfico de este banco de filtros (véase [46]).

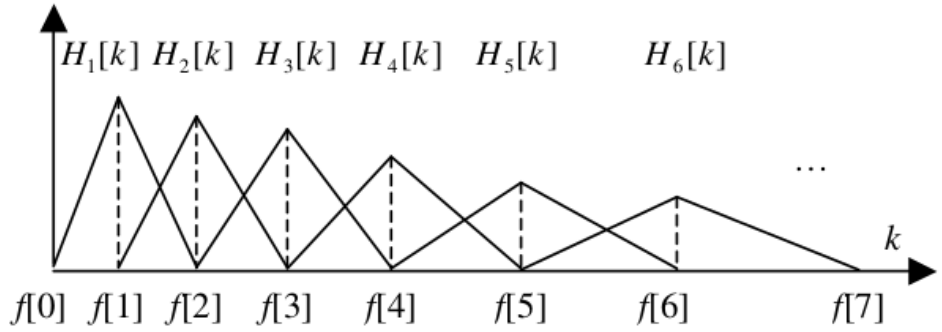


Figura 3.5: Banco de filtros triangulares. Fuente: [46]

A partir de estos filtros, el re-escalamiento de las frecuencias en la escala de Mel y la transformada de coseno discreto se construyen M valores conocidos como los coeficientes cepstrales (denotados como $c[n]$):

$$c[n] = \sum_{m=0}^{M} S[m] cos\left(\pi n \frac{m + \frac{1}{2}}{M}\right) \quad 0 \leq n < M, \tag{3.13}$$

donde,

- $S[m]$ es la expresión:

$$S[m] = \ln\left(\sum_{k=0}^{N-1} |X[k]|^2 H_m[k]\right) \quad 0 \leq m < M. \tag{3.14}$$

- $|X[k]|^2$ es la norma euclidiana al cuadrado de la transformada de Fourier discreta de la señal x.
- $H_{,}$ es el $m-$ésimo filtro triangular evaluado en k.

3.2.6. Normalización del Vector de Características

La normalización **min-max** (véase [52]) es un método comúnmente utilizado para normalizar un vector de características, y se define matemáticamente de la siguiente manera:

Para un vector de características $X = [X_1, X_2, ..., X_n]$, la normalización min-max produce un vector normalizado y denotado por $Y = [Y_1, Y_2, ..., Y_n]$ como:

$$Y_i = \frac{(X_i - minX_i)}{(max(X_i) - min(X_i))}, \tag{3.15}$$

donde $min(X_i)$ y $max(X_i)$ son el valor mínimo y máximo en el conjunto de características del que proviene cada entrada de X.

3.2.7. Winsorización de las Características

Winsorizar los datos es una técnica de preprocesamiento de datos que se utiliza para reducir la influencia de los valores atípicos (*outliers*) en una distribución de datos (véase [51]). Esta técnica consiste en reemplazar los valores atípicos en una distribución con un valor máximo o mínimo que se define previamente, en lugar de eliminarlos por completo. La formulación matemática de la técnica de winsorización es la siguiente:
Sea X un conjunto de datos y k un número real que se utiliza para definir el porcentaje de valores atípicos que se desea eliminar (usualmente, $k \in (0,0.5)$. Entonces, la técnica de winsorización se puede definir como:

1. Ordenar los valores de X de forma ascendente, y calcular los percentiles k y $1-k$ de la distribución. Estos percentiles representan los valores en X que se encuentran en el percentil k y en el percentil $1-k$, respectivamente.
2. Calcular el rango intercuartil (IQR) de X como la diferencia entre el percentil $1-k$ y el percentil k de la distribución, i.e.

$$IQR_w := Q_{1-k} - Q_k\} \quad \textbf{Rango intercuartil de winsor.} \tag{3.16}$$

   donde Q_{75} y Q_{25} son los percentiles 75 y 25, respectivamente.
3. Calcular los límites superior e inferior de la técnica de winsorización, definidos como

$$min_w := Q_k - 1{,}5 * IQR_w\} \quad \textbf{Límite inferior de winsor.} \tag{3.17}$$

$$max_w := Q_{1-k} + 1{,}5 * IQR_w\} \quad \textbf{Límite superior de winsor.} \tag{3.18}$$

   donde Q_k y Q_{1-k} son los percentiles k y $1-k$ de X, respectivamente, e IQR es el rango intercuartil de X.
4. Reemplazar todos los valores en X que son menores que el límite inferior de winsorización con el valor del límite inferior. De manera similar, reemplazar todos los valores en X que son mayores que el límite superior con el valor del límite superior.
5. Calcular los valores estadísticos de interés (por ejemplo, la media o la varianza) de la distribución winsorizada resultante.

3.3. Métricas

3.3.1. Métricas para Reconocimiento de Género

- Se definen las categorías a trabajar: Hombres y Mujeres.
- A cada categoría se les asigna una variable distinta:
  - P:= hombre.
  - N:= mujer.
- Posteriormente, se definen los 4 escenarios en el problema de clasificación binaria:
  - TP:= Predicción hombre y realidad hombre.
  - FP:= Predicción hombre y realidad mujer.

- TN:= Predición mujer y realidad mujer.
- FN:= Predicción mujer y realidad hombre.

- Así, se define los siguientes subconjuntos con sus elementos.
  - 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 hombres.
  - FN + TN := Total de predicciones de mujeres.
  - TP + FN := Total de hombres.
  - FP + TN := Total de mujeres.
- Las fórmulas de las métricas que se utilizarán para evaluar los modelos son las siguientes:

$$\text{Exactitud } (\textit{Accuracy}) := \frac{TP + TN}{TP + TN + FP + FN} \tag{3.19}$$

$$\text{Precisión } (\textit{Precision}) := \frac{TP}{TP + FP} \tag{3.20}$$

$$\text{Exhaustividad } (\textit{Recall}) := \frac{TP}{TP + FN} \tag{3.21}$$

$$\text{Puntaje F1 } (\textit{F1-score}) := \frac{2\times\text{Precisión}\times\text{Exhaustividad}}{\text{Precisión}+\text{Exhaustividad}} \tag{3.22}$$

donde:

- La exactitud es la proporción de predicciones correctas sobre todas las predicciones.
- La precisión es la proporción de predicciones correctas de hombres de todo el conjunto de predicciones de hombres.
- La exhaustividad es la proporción de predicciones correctas de hombres entre todo el conjunto de hombres.
- El puntaje F1 es la media armónica entre la exhaustividad y la precisión. Es usada cuando la cantidad de instancias en las categorías no están equilibradas.
- Como última métrica, se considera el área bajo la curva ROC. La curva ROC se construye trazando la tasa de verdaderos positivos (*True Positive Rate* o *TPR* por sus siglas en inglés) en el eje y, y la tasa de falsos positivos (*False Positive Rate* o *FPR*) en el eje x, mientras se varía el umbral de decisión del modelo. Es decir dicha curva está compuesta por los puntos:

$$(X, Y) = (FPR, TPR) = \left(\frac{FP}{FP + TN}, \frac{TP}{TP + FN}\right) \tag{3.23}$$

  La curva muestra la relación entre la sensibilidad y la especificidad del modelo a diferentes umbrales de decisión. El área bajo la curva ROC es una medida cuantitativa de la capacidad del modelo para distinguir entre las dos clases. El valor del AUC-ROC varía entre 0 y 1, donde un valor de 0.5 indica que el modelo es tan bueno como un clasificador aleatorio, y un valor de 1.0 indica que el modelo puede distinguir perfectamente entre las dos clases.

3.3.2. Métricas para Reconocimiento de Género y Edad

Las métricas utilizadas se describen a continuación:
Considere que hay n clases de interés, se define como verdaderos positivos de clase i (denotado con TP_{ii}) como el número de ejemplos de la clase i que fueron clasificados correctamente en la clase i. Luego, se definen los falsos positivos de la clase i en la clase j (denotado como FP_{ij}) como el número de ejemplos de la clase i

que fueron clasificados en la clase j. En la tabla 3.1, se muestra un ejemplo de matriz de confusión para N clases.

Tabla 3.1: Matriz de confusión para N clases

Clases reales / Clases predichas	**Clase 1**	**Clase 2**	...	**Clase N**
Clase 1	TP_{11}	FP_{12}	...	FP_{1N}
Clase 2	FP_{21}	TP_{22}	...	FP_{2N}
...	...	...	...	...
Clase N	FP_{N1}	FP_{N2}	...	TP_{NN}

Por lo anterior, La precisión *(precision)* se define como la proporción de ejemplos clasificados como la clase i que realmente pertenecen a la clase i:

$$precision_i = \frac{TP_{ii}}{\sum_{j=1,\ j\neq i}^{N} FP_{ij} + TP_{ii}} \tag{3.24}$$

La exhaustividad (*recall*) se define como la proporción de ejemplos de la clase i que fueron correctamente clasificados:

$$recall_i = \frac{TP_{ii}}{TP_{ii} + \sum_{i=1,\ i\neq j}^{n} FP_{ij}} \tag{3.25}$$

La exactitud (*accuracy*) se define como la proporción de ejemplos clasificados correctamente en todas las clases:

$$accuracy = \frac{\sum_{i=1}^{N}(TP_{ii})}{\sum_{i=1}^{N}(TP_{ii}) + \sum_i \sum_j FP_{ij}} \tag{3.26}$$

La medida F_1 (F_1-*Measure*) es una media armónica entre *precision* y *recall*, que se utiliza para obtener una medida única de la calidad de la clasificación:

$$F_1 score_i = \frac{2 * precision_i * recall_i}{precision_i + recall_i} \tag{3.27}$$

4 | Análisis, Parámetros y Diseño de los Sistemas

En este capítulo se exponen las herramientas de software utilizadas, así como los parámetros empleados para los procesos de preprocesamiento y extracción de características. De manera general los puntos a tratar son los siguientes:

- PRAAT,*Parselmouth PRAAT*, Keras y WEKA.
- Preprocesamiento de los audios (Substracción espectral y filtro detector de voz).
- Extracción de características acústicas y cepstrales.
- Desglose de las arquitecturas de las redes neuronales.

4.1. Librerías y Softwares Utilizados

- PRAAT (véase [53]) es un software de código abierto utilizado para el análisis acústico y fonético de señales de audio. Fue desarrollado por Paul Boersma y David Weenink en la Universidad de Ámsterdam en la década de 1990 y se ha convertido en una herramienta de referencia en la investigación en lingüística, fonética y ciencias del habla. PRAAT permite visualizar y manipular señales de audio, analizar su espectro, calcular medidas acústicas como la frecuencia fundamental, la intensidad y la duración, y generar visualizaciones de los datos en forma de gráficos y espectrogramas.
- Parselmouth PRAAT [56] es una biblioteca de Python que proporciona una interfaz para PRAAT a través del lenguaje de programación Python. Esta biblioteca permite a los usuarios acceder a las funciones de Praat desde Python, lo que facilita la automatización de tareas de análisis de audio y la integración con otros paquetes de Python para el procesamiento de datos. Parselmouth PRAAT también proporciona una serie de herramientas adicionales para el análisis de audio, como la extracción de características acústicas de señales de audio, el análisis de la voz y la identificación de patrones de habla.
- Keras es una biblioteca de código abierto para el aprendizaje profundo (*Deep Learning*) escrita en el lenguaje de programación Python. Fue desarrollada por François Chollet en 2015 y ha ganado popularidad en la comunidad de la inteligencia artificial debido a su facilidad de uso y flexibilidad. Keras está diseñada para ser una interfaz de alto nivel para la construcción de redes neuronales y permite a los usuarios desarrollar modelos de aprendizaje profundo de manera eficiente y sin necesidad de conocimientos profundos de matemáticas o programación.
- WEKA consiste en una compilación de métodos de aprendizaje automático destinados a la exploración de datos, ofreciendo recursos para manipular datos, categorizar, predecir, organizar, descubrir patrones y representar gráficamente resultados.

4.2. Parámetros Relacionados al Preprocesamiento de los Audios

Los parámetros utilizados para la substración espectral en el software PRAAT fueron:

- Rango del tiempo con ruido (*Noise time range (s)*): 0.0, 0.0. (valores predefinidos).
- Longitud de ventana con traslape (*Window length (s)*: 0.025 segundos.
- Rango de frequencias de filtro (*Filter frequency range (Hz)*:) 80 Hz a 10000.0 Hz.
- Suavización (*Smoothing (Hz)*: 40 Hz.
- Método de reducción de ruido: *Spectral substraction*

El algoritmo de detección de actividad de voz implementado en este trabajo se resume en los siguientes pasos:

1. Seleccionar la señal de voz de entrada y obtener su duración total.
2. Aplicar un script de preprocesamiento para preparar la señal para su análisis.

3. Calcular los límites mínimo y máximo de la frecuencia fundamental de la señal utilizando la función *minmaxf0.praat*.
4. Convertir la señal en un punto de proceso (point process) utilizando los límites calculados en el paso anterior.
5. Crear un marcador *TextGrid* para la señal de voz utilizando los límites del punto de proceso.
6. Iterar sobre cada intervalo del *TextGrid* y determinar si es una región sonora (denotada con V) o no sonora (denotada con U).
7. Para cada región sonora, eliminar las partes de la señal que no corresponden a la voz sonora.
8. Para cada región no sonora, suavizar las transiciones de la señal para eliminar los clics de inicio y finalización.
9. Aplicar un script de postprocesamiento para preparar la señal para su exportación.
10. Renombrar las señales de voz sonora y no sonora.
11. Si se especifica, crear un nuevo *TextGrid* para la señal de voz utilizando los intervalos VUV (*voiced/unvoiced*) y exportar la señal.
12. Eliminar los objetos creados en el proceso.
13. Finalmente, se reagrupan únicamente los audios que se detectaron como sonoros con un traslape de 0.1 segundos.

La aplicación de estos dos filtros (substracción espectral y detector de actividad de voz) en un audio puede visualizarse en la Figura 4.1.

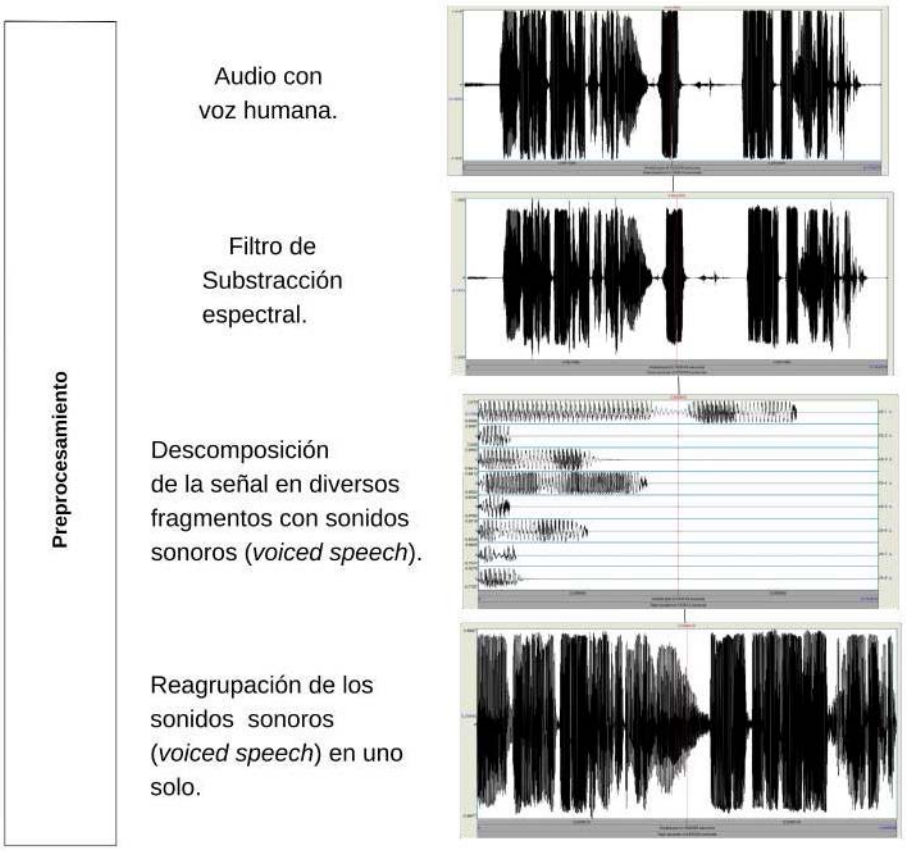


Figura 4.1: Proceso de atenuación de ruido mediante substracción espectral y filtro de actividad de voz.

4.3. Características Extraídas

Una manera de visualizar los grupos de características extraídas (tono, formantes e intensidad) se muestra en la Figura 4.2. Esta figura muestra una ventana del software PRAAT en la que se muestra la señal (voz) en el dominio del tiempo, mientras que en el dominio de frecuencias se visualizan los contornos de los formantes, la frequencia funtamental y la intensidad de la voz. Los parámetros establecidos para obtenerlos se exponen a continuación.

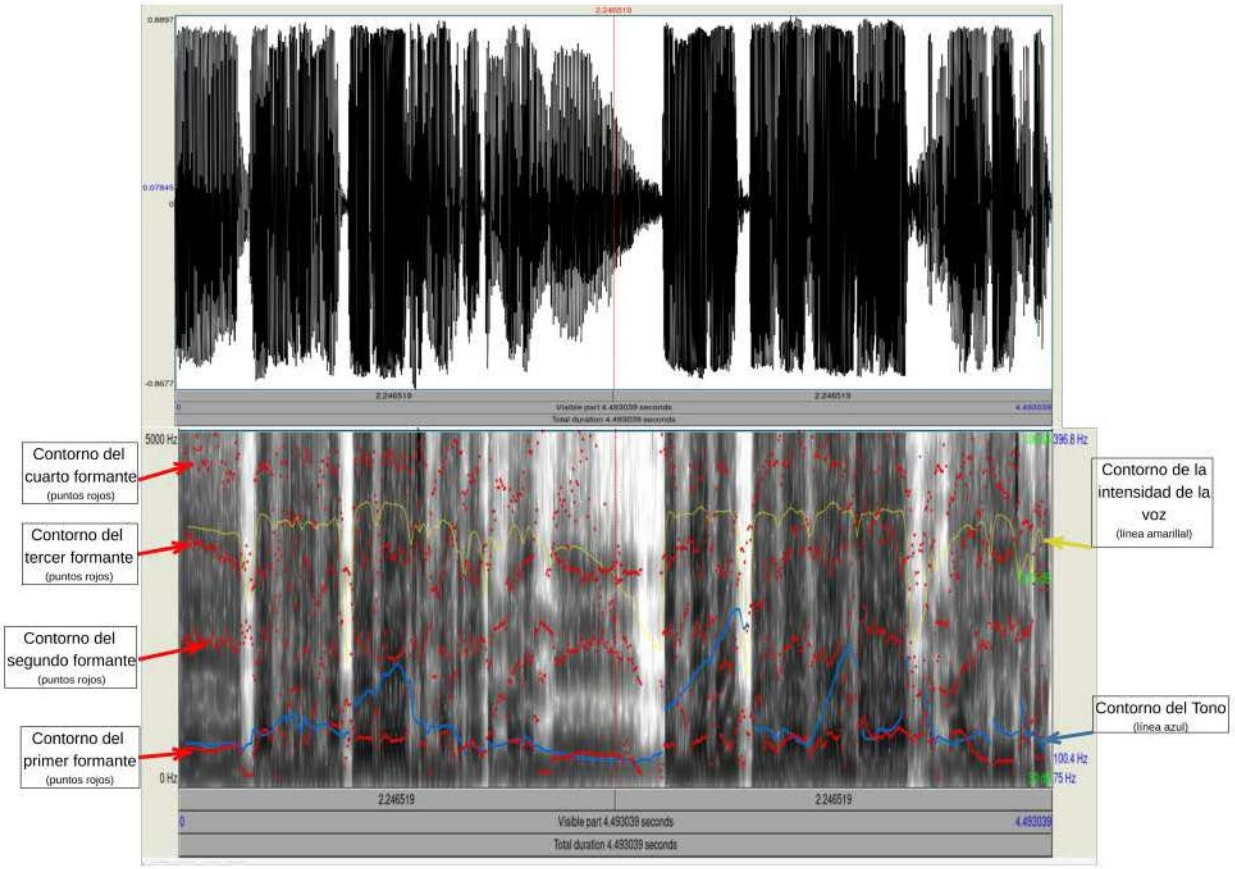


Figura 4.2: Características acústicas (tonales, formantes e intensidad) mostradas en un audio traslapado.

4.3.1. Características Derivadas de la Frecuencia Fundamental

Las características estadísticas derivadas de la frecuencia fundamental empleadas fueron

- Mínimo valor del tono.
- Primer cuartil del tono.
- Media del tono.
- Mediana del tono.
- Tercer cuartil del tono.
- Máximo valor del tono.
- Desviación estándar del tono.

Los parámetros utilizados para su extracción fueron los siguientes:

- Paso del tiempo (*Time step (s)*): auto.
- Tono mínimo (*Pitch floor (Hz)*: 75.0

- Máximo número de candidatos (*Max. number of candidates*): 15
- Umbral de silencio (*Silence threshold):* 0.03
- Umbral de voz (*Voicing threshold*): 0.45
- Costo de octava (*Octave cost*): 0.35
- Costo sonoro/no sonoro (*Voiced / Unvoiced cost*): 0.14
- Tono máximo (*Pitch ceiling*): 600 / 350[1]

4.3.2. Características Derivadas de los Formantes

Se extrajo la mediana de los valores de los cuatro primeros formantes por separado. Los parámetros utilizados en la implementación se desglosan a continuación:

- Paso de tiempo (s)(*Time step (s)*): el tiempo entre los centros de fotogramas de análisis consecutivos. Se introdujo la opción auto.
- Número máximo de formantes (*Max. number of formants:)* 5.
- Techo de formantes (Hz) (*Maximun Formant)*: La frecuencia máxima del rango de búsqueda de formantes, en hercios. Se ajustó a 8000 Hz.
- Duración de la ventana (s) (*Window length (s)*: Duración efectiva de la ventana de análisis, en segundos. Se especificó en 25 milisegundos.
- Preénfasis de (Hz) (*Pre-emphasis from (Hz)*: El punto de +3 dB para un filtro pasa bajo invertido con una pendiente de +6 dB/octava. Se indicó en 50 Hertz.
- Número de desviaciones estándar (*Number of std. dev.*: 1.5
- Máximo número de iteraciones (*Maximum number of iterations:*) 5
- Tolerancia (*Tolerance*): 0. 0000001

4.3.3. Características Derivadas de la Intensidad de la Voz

Se extraen el primer cuartil, la media, la mediana y el tercer cuartil de la intensidad de voz calculada. Los parámetros empleados fueron:

- Tono mínimo (*Minimum pitch (Hz)*) 75 Hertz.
- Tiempo (*Time step (s)*) 0.0 (auto).
- Substracción del promedio (*Substract mean) ON.*

4.3.4. Características Derivadas del Espacio Cepstral

Se seleccionan los coeficientes cepstrales más relevantes para la representación de la señal de sonido. Para este trabajo se utilizaron 30 coeficientes. Posteriormente se calcula la mediana de esas 30 series de coeficientes. Los primeros treinta coeficientes MFCC se pueden interpretar de la siguiente manera ([81]):

- MFCC1: se asocia con la energía global de la señal de habla y se utiliza comúnmente para detectar cambios de intensidad y para normalización de la señal.
- MFCC2: se relaciona con la información de tono y entonación de la señal de habla, y se utiliza para la detección de la modulación de la frecuencia fundamental.
- MFCC3 a MFCC7: se asocian con la forma del tracto vocal y los armónicos de la señal de habla. Estos coeficientes se utilizan a menudo para identificar y distinguir diferentes consonantes y vocales.
- MFCC8 a MFCC14: se relacionan con la envolvente de la señal de habla y su variación en el tiempo. Estos coeficientes se utilizan a menudo para la identificación de los diferentes hablantes y para la detección de cambios de entonación.

[1]Las primeras dos metodologías para reconocimiento de género emplearon el valor de 600 Hertz, las demás metodologías utilizaron el valor de 350 Hertz.

- MFCC15 a MFCC22: se relacionan con las características espectrales de alta frecuencia de la señal de habla, como las formantes de alta frecuencia y el ruido de banda ancha.
- MFCC23 a MFCC30: se relacionan con las características espectrales de muy alta frecuencia de la señal de habla, como los armónicos de alta frecuencia y el ruido de banda estrecha.

Los parámetros utilizaron para la implementación de PRAAT fueron:

- Número de coeficientes *Number of coefficients*: 30
- Longitud de ventana (*Window length (s):* 0.015 s
- Salto de tiempo (*Time step (s)*): 0.05s
- Parámetros para los bancos de filtros.
  - Primer filtro de mel (*First filter frequency (mel)*): 100.0
  - Distancia entre filtros (*Distance between filters (mel)*): 100.0
  - Frequencia máxima de mel (*Maximum frequency (mel)*): 0.0

4.4. Arquitecturas de las Redes

El desglose de las arquitecturas de las redes diseñadas para el reconocimiento de género y edad pueden verse en la Figura 1.3. Se inicia con una breve descripción de las redes tipo perceptrones multicapa.
El algoritmo de perceptrón multicapa (MLP, por sus siglas en inglés) es un tipo de red neuronal artificial que se utiliza para clasificación y regresión. Se compone de varias capas, incluyendo una capa de entrada, una o varias capas ocultas y una capa de salida. Cada capa consta de un conjunto de neuronas, y cada neurona se conecta a todas las neuronas de la capa siguiente. El MLP utiliza la retropropagación del error para ajustar los pesos de las conexiones y aprender de los datos de entrenamiento. Las arquitecturas empleadas pueden verse en las Figuras 4.3 y 4.4. En términos generales se contempla:
Red detectora de género y edad (adultos y adolescentes, véase Fig. 4.3 lado izquierdo):

1. Cuatro capas densas con 512, 256, 128 y 128, respectivamente. Cada neurona tiene la función de activación *ReLU (Rectified Linear Unit)*. Su fórmula matemática es:

$$f(x) = max\{0, x\} \tag{4.1}$$

2. Una capa de dilución (*DropOut*) de $p = 10\,\%$, que es una técnica de regularización que se utiliza en los MLP para evitar el sobreajuste. Durante el entrenamiento, se apaga aleatoriamente un porcentaje de las neuronas en la capa anterior (es decir, la capa oculta anterior). La fórmula matemática para la capa *DropOut* es:

$$y_i = r_i * x_i, \tag{4.2}$$

   donde x_i es la entrada de la neurona i en la capa anterior, r_i es un valor binario aleatorio que es 1 con una probabilidad p y 0 con una probabilidad $1 - p$, y y_i es la salida de la neurona i en la capa *DropOut*.
3. Una capa densa con 64 neuronas y función de activación *ReLU*, seguida de una capa *Dropout* con $p = 10\,\%$.
4. Finalmente, una capa externa de cuatro neuronas, con función de activación $Softmax$ (véase [71]). Esta función, se utiliza comúnmente en la capa de salida de los MLP para problemas de clasificación multiclase. Su fórmula matemática es:

$$Softmax(X_i) = \frac{e^{X_i}}{\sum_{j=1}^{n} e^{X_j}}, \tag{4.3}$$

   donde X_i es la entrada de la neurona i en la capa de salida, e es la función exponencial y se realiza una suma de todas las entradas exponenciales en la capa de salida. La función *Softmax* produce una distribución de probabilidad en la salida, es decir, las salidas de la capa de salida suman 1. La salida con mayor probabilidad es la que se elige como resultado.

5. Finalmente, la función de pérdida fue *sparse categorical crossentropy* que se utiliza comúnmente en modelos de aprendizaje profundo para la clasificación multiclase en la que las etiquetas son enteros en lugar de codificaciones *one-shot*. Esta función de pérdida se define matemáticamente como:

$$L(y, \acute{y}) = -\sum_i y_i log(\hat{y}_i) \tag{4.4}$$

donde y es el vector de etiquetas verdaderas (un tensor de forma (*batch size*[2],) que contiene enteros que representan las etiquetas de clase para cada ejemplo en el lote), y $\hat{y}$ es el vector de salidas predichas por el modelo (un tensor de forma (*batch size*, *num classes*[3]) que contiene las salidas predichas para cada clase para cada ejemplo en el lote). La función de pérdida calcula la suma de las pérdidas logarítmicas negativas para cada clase para cada ejemplo en el lote.

Red detectora de edad en décadas (véase Fig. 4.3 lado derecho):

1. Una capa densa de 1024 neuronas con función de activación *ReLU*.
2. Una capa de dilución con $p = 10\,\%$.
3. Una capa densa de 2048 neuronas con función de activación *ReLU*.
4. Finalmente, una capa externa de seis neuronas con función de activación *SoftMax*.

Red detectora de género y edad en décadas (véase Fig. 4.4):

1. Una capa de 512 neuronas con función de activación *ReLU*.
2. Una capa de dilución con $p = 10\,\%$.
3. Dos capas densas con 512 neuronas y función de activación *ReLU*.
4. Una capa de dilución con $p = 10\,\%$.
5. Dos capas densas con 512 neuronas y función de activación *ReLU*.
6. Una capa de dilución con $p = 10\,\%$.
7. Una capa de salida con 12 neuronas con función de activación *SoftMax*.

[2]El tamaño del lote (o *batch size*), es el número de ejemplos que se procesan antes de que se actualice el modelo con los gradientes calculados a partir de ese lote.

[3]Es el número de clases a identificar.

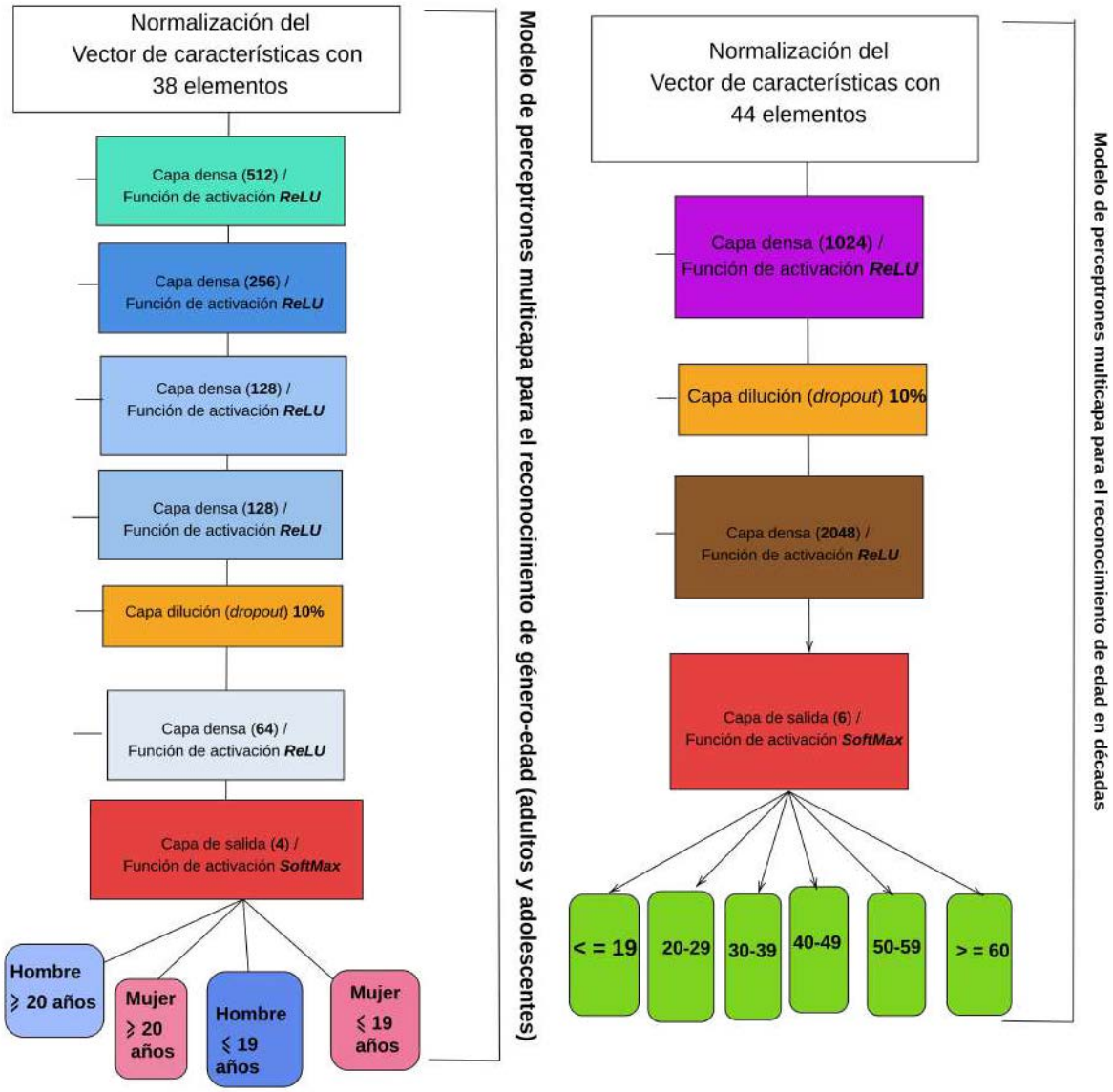


Figura 4.3: Arquitectura de la red de perceptrones multicapa detectora de género y edad (adultos y adolescentes, derecha) y edad en décadas (izquierda).

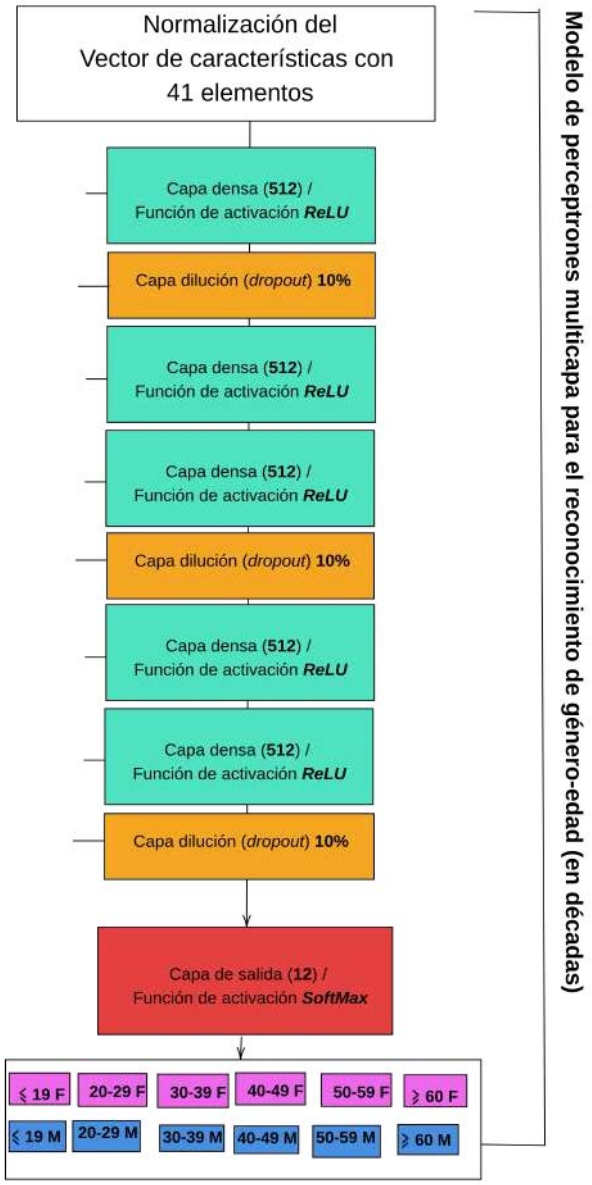


Figura 4.4: Arquitectura de la red de perceptrones multicapa detectora de género y edad (en décadas). La letra F denota al género femenino, mientras que la M denota al género masculino.

La metodología para validar los modelos anteriores se encuentra detallada en el siguiente capítulo. No obstante, de manera general, cada uno de los modelos generados anteriormente se pondrá a prueba en diferentes conjuntos de datos. Estos deben tener las mismas características con las que fueron entrenados.

5 | Experimentación y Resultados

Los casos de experimentación que se muestran a continuación son seis, uno por cada metodología propuesta. Para facilitar la lectura de esta sección, cada metodología se presenta con tres apartados:

- **Introducción**: Información preliminar del contexto de la metodología.
- **Objetivo**: Exposición del objetivos general y los objetivos específicos de la metodología.
- **Características de las pruebas realizadas**: Se exponen el tipo de conjunto de voces utilizados y su información estadística.
- **Resultados**: Se presentan los resultados de la metodología.

Tras exponer cada uno de los resultados de las metodologías, se tendrá una discusión general de sus desempeños en relación al problema de reconocimiento de género y edad por voz.

5.1. Primer Caso de Prueba del Reconocimento de Género por Voz

- **Introducción**
  Una forma de distinguir el género de una voz es a través del dimorfismo sexual, es decir, las variaciones en la fisionomia entre hombres y mujeres. En cuestiones acústicas, la frecuencia fundamental (F_0) de una voz masculina tiende a ser menor que su contraparte femenina (véase [50]). Si bien la edad y el idioma puede afectar los rangos de detección para hombres y mujeres, se requiere de estudios que permitan validar un grupo de características robustas ante múltiples idiomas.
- **Objetivo**
  El objetivo de esta metodología es analizar el desempeño de las características estadísticas derivadas de la frecuencia fundamental en el reconocimiento de género.
  Los objetivos específicos son
  - Realizar el reconocimiento de género en los idiomas Inglés, Español, Alemán, Francés, Chino y su combinación por medio de las características estadísticas mínimo de F_0, primer cuartil de F_0, media de F_0, mediana de F_0, tercer cuartil de F_0, Máximo de F_0, desviación estándar de F_0, Rango intercuartil de F_0 y la estimación de la edad en décadas.
  - Validar las estadísticas empleadas mediante el estudio de escenarios balanceados y no balanceados donde cada uno tendrá los casos de voces repetidas y voces únicas.
  - Emplear tres tipos diferentes de clasificadores (Perceptrón multicapa, Árbol de búsqueda J48, y regresión lógística) para comparar las métricas obtenidas.

  La primera metodología denominada sistema base (véase Figura 5.1) consiste en los siguientes pasos:
  1 Para cada uno de los audios de la base de datos de *Mozilla Common voice* en un idioma específico se realiza lo siguiente.
    1.1 Aplicar un filtro de detección de voz y calcular la serie tonos (también conocido como contorno de tono) en un rango de 75 a 600 Hertz. Dicho rango fue propuesto por el Dr. Paul Boersma para analizar las voces de ambos géneros [62].
    1.2 A cada serie de tono se le extrajeron 8 características estadísticas y la etiqueta de edad (en décadas).
  2 Se construyen los escenarios balanceados (misma cantidad de audios por género) y no balanceados. Para cada escenario se divide en voces únicas (un audio por hablante) y voces repetidas (varios audios por hablante).
  3 Se configura el software WEKA en la validación cruzada *ten-cross fold validation*.
  4 Se entrenan los algoritmos: Perceptrón multicapa, árbol de búsqueda J48 y regresión logística con los parámetros establecidos en WEKA.

5 Se evaluan las clasificaciones realizadas en términos de sus métricas: precisión, exhaustividad, puntaje F1 y área bajo la curva ROC.

Cabe destacar que no se utilizaron filtros para atenuación del ruido proveniente del ambiente de grabación. Si bien esto ocasionó que varios de los valores estimados de F_0 en los audios tuvieran una mayor presencia de ruido, la representación mediante los estadísticos robustos (como la mediana) reflejan una diferencia estadística significativa en los géneros (véase las tablas 5.2, 5.3, 5.4, 5.5, 5.6 y 5.7).

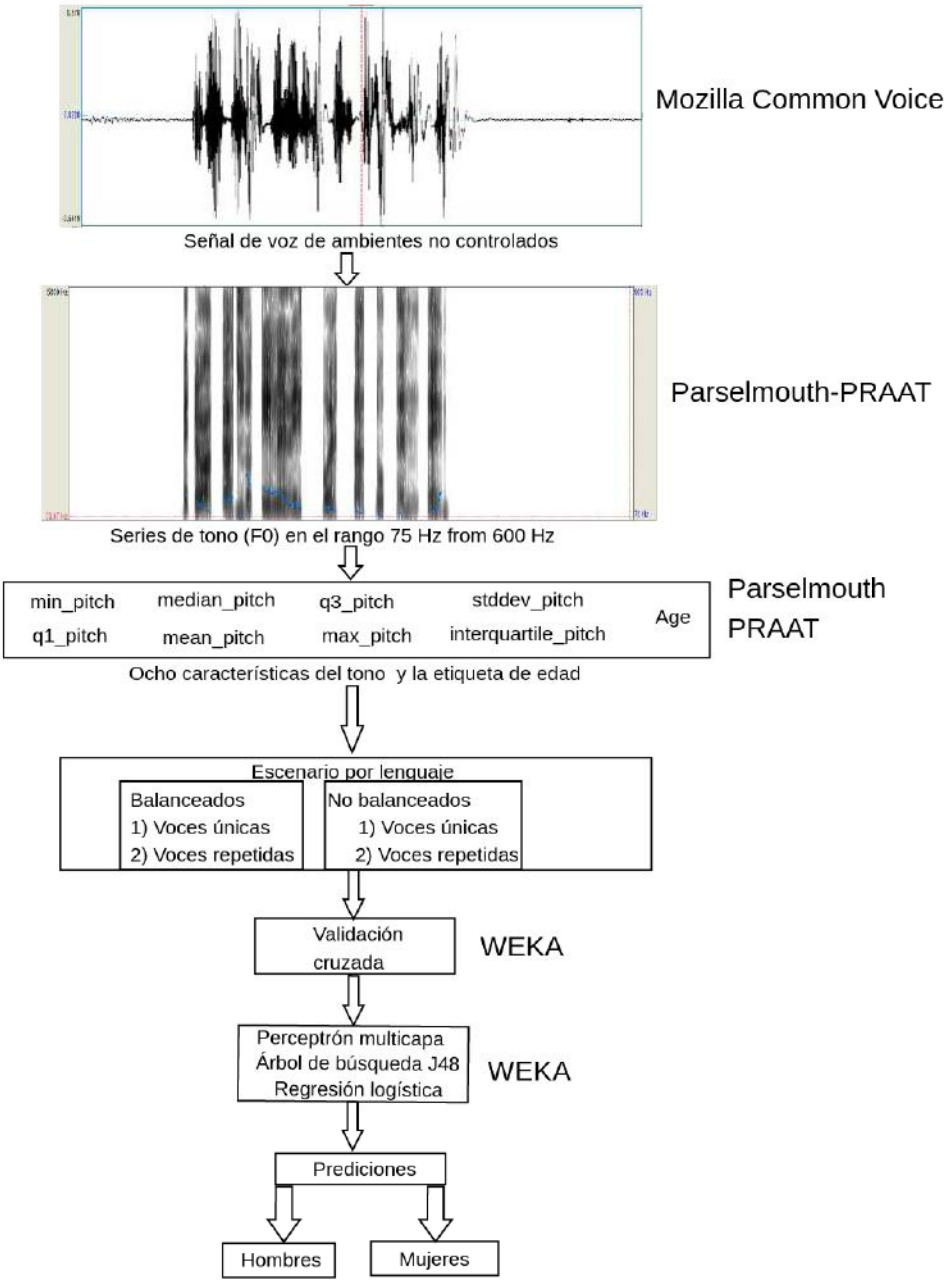


Figura 5.1: Metodología 1 propuesta para el reconocimiento de género por voz mediante algoritmos de aprendizaje clásicos.

- **Características de las Pruebas Realizadas**
  Los audios fueron tomados del corpus *Mozilla Common Voice* [13] en los idiomas Español, Alemán, Francés, Inglés y Chino. Cada conjunto de voces se constituye por un conjunto de entrenamiento, uno de prueba y uno de validación. Para este experimento, se combinaron estos tres conjuntos. Además, solo se consideraron las personas cuyo rango de edad estuviera entre los 20 años hasta los 60 años. Posteriormente, se construyeron los escenarios (véase la Tabla 5.1) Balanceados (misma cantidad de audios en cada clase) y no balanceados. Estos fueron divididos en dos casos: Voces repetidas (más de un audio por hablante) y voces únicas (un audio por hablante). Se configuró de esta manera, ya que un conjunto de voces en donde cada audio es proveniente de diferentes personas, provee una mayor variabilidad en sus características, que uno con multiplicidad de audios.

Tabla 5.1: Resumen tabular de escenarios balanceados y no balanceados.

Balanceados	**No balanceados**
Voces repetidas (audios por género)	**Voces repetidas**
Español (33,351 audios).	Español (210,189 hombres y 33,351 mujeres).
Alemán (45,081 audios).	Alemán (417,178 hombres y 45,081 mujeres).
Francés (32,652 audios).	Francés (303,351 hombres y 32,652 mujeres).
Inglés (168,064 audios).	Inglés (528,029 hombres y 168,064 mujeres).
Chino (11,391 audios)	Chino (34,958 hombres y 11,391 mujeres).
Voces únicas (audios por género)	**Voces únicas**
Español (791 audios).	Español (2576 hombres y 791 mujeres).
Alemán (433 audios).	Alemán (2815 hombres y 433 mujeres).
Francés (499 audios).	Francés (2531 hombres y 499 mujeres).
Inglés (2377 audios).	Inglés (8867 hombres y 2377 mujeres).
Chino (182 audios).	Chino (714 hombres y 182 mujeres).

Tabla 5.2: Descripción de la mediana de la frecuencia fundamental en grupos de edades en el idioma español.

Edad	**Género**	**Cantidad**	**Media**	**Desv. est.**	**Mínimo**	**Primer cuartil**	**Mediana**	**Tercer cuartil**	**Máximo**
Veinte	**Mujer**	15190.0	**216.032805**	29.143248	79.348518	197.871444	**216.627463**	232.926343	550.883335
	Hombre	55066.0	**131.236083**	31.148195	76.442399	116.242104	**127.836973**	140.778847	588.114830
Treinta	**Mujer**	7754.0	**200.043885**	32.864939	87.906049	183.928191	**199.795923**	217.212625	572.116091
	Hombre	35686.0	**123.271570**	30.569502	77.858920	106.625494	**119.982865**	134.902871	594.813063
Cuarenta	**Mujer**	6128.0	**199.906734**	38.953387	82.169328	179.309078	**194.404718**	215.324638	595.367925
	Hombre	18832.0	**126.027095**	46.017092	76.381957	108.914127	**119.136981**	130.706707	569.874027
Cincuenta	**Mujer**	3268.0	**182.221728**	38.888553	94.582082	157.576701	**179.495143**	209.668450	564.047397
	Hombre	27883.0	**109.138906**	25.411245	78.808535	97.830621	**105.424912**	114.460911	595.737802
Sesenta	**Mujer**	1011.0	**195.232375**	30.808213	78.051816	184.565525	**199.230044**	213.921323	274.220553
	Hombre	72722.0	**137.734534**	39.202579	78.923668	124.716757	**131.848230**	140.969347	590.433918

Tabla 5.3: Descripción de la mediana de la frecuencia fundamental en grupos de edades en el idioma alemán.

Edad	Género	Cantidad	Media	Desv. est.	Mínimo	Primer cuartil	Mediana	Tercer cuartil	Máximo
Veinte	**Mujer**	13384.0	**217.726975**	34.121543	81.330732	198.882410	**219.260043**	237.738045	571.533359
	Hombre	126448.0	**133.789340**	43.314516	75.142126	112.873344	**125.222813**	143.222707	594.112205
Treinta	**Mujer**	7451.0	**194.784597**	30.068539	98.063295	178.358411	**191.686355**	206.277973	556.866022
	Hombre	98211.0	**128.668474**	42.371934	76.851089	110.060108	**121.449512**	135.865032	595.531619
Cuarenta	**Mujer**	6317.0	**195.951111**	22.688597	120.030989	181.575969	**193.685621**	208.165086	396.412418
	Hombre	117302.0	**115.486767**	43.653495	75.440183	95.010048	**106.327603**	125.141279	593.074262
Cincuenta	**Mujer**	12506.0	**195.014688**	25.198110	91.250160	178.666424	**197.570416**	211.333617	466.041805
	Hombre	66435.0	**117.252529**	34.811391	76.562509	99.761109	**109.134911**	126.558792	588.254103
Sesenta	**Mujer**	5423.0	**208.950091**	28.215696	115.879691	184.664755	**209.095419**	233.190191	330.316530
	Hombre	8782.0	**139.486134**	49.868393	78.693117	116.394428	**131.716312**	148.481550	586.599943

Tabla 5.4: Descripción de la mediana de la frecuencia fundamental en grupos de edades en el idioma francés.

Edad	Género	Cantidad	Media	Desv. est.	Mínimo	Primer cuartil	Mediana	Tercer cuartil	Máximo
Veinte	**Mujer**	13497.0	**214.325347**	36.337171	80.282027	197.202183	**217.948827**	237.312728	496.759279
	Hombre	86982.0	**126.668270**	37.029414	76.181364	108.986242	**121.106608**	135.195934	597.071396
Treinta	**Mujer**	6486.0	**213.486106**	28.273379	124.822570	195.435364	**210.603677**	225.870261	564.869153
	Hombre	85594.0	**125.747242**	36.491598	75.934522	111.640797	**121.457782**	131.633993	593.964602
Cuarenta	**Mujer**	5745.0	**205.452866**	24.276347	133.354946	191.947946	**203.167448**	217.902257	499.783365
	Hombre	73420.0	**131.505942**	38.169515	75.739288	111.010758	**126.682466**	143.409480	598.393021
Cincuenta	**Mujer**	4530.0	**203.242812**	25.535449	101.542996	192.547038	**203.690754**	215.487576	552.954671
	Hombre	47989.0	**113.248773**	29.340315	76.213059	96.929860	**106.578577**	121.511811	595.498197
Sesenta	**Mujer**	2394.0	**197.404239**	30.197854	119.613900	174.089558	**194.379064**	220.922567	359.290722
	Hombre	9366.0	**117.457865**	26.263285	79.525602	102.216299	**111.716514**	125.942964	552.396635

Tabla 5.5: Descripción de la mediana de la frecuencia fundamental en grupos de edades en el idioma inglés.

Edad	Género	Cantidad	Media	Desv. est.	Mínimo	Primer cuartil	Mediana	Tercer cuartil	Máximo
Veinte	**Mujer**	74115.0	**220.692424**	33.506800	81.992445	198.670091	**218.467374**	239.848024	566.282901
	Hombre	199630.0	**131.831001**	41.968473	76.622973	110.458580	**123.302696**	141.672409	588.641699
Treinta	**Mujer**	26833.0	**200.331971**	32.387608	82.654962	181.399641	**198.373820**	216.334192	560.146522
	Hombre	144331.0	**126.177323**	43.766064	75.292111	106.997729	**119.003597**	132.563695	597.414785
Cuarenta	**Mujer**	20777.0	**203.590394**	45.562661	82.067515	170.130769	**202.213418**	241.432601	559.325092
	Hombre	108591.0	**117.077327**	43.520292	75.895589	93.477505	**108.057051**	127.747548	597.053249
Cincuenta	**Mujer**	25120.0	**201.208526**	45.019467	84.078296	168.196034	**194.435207**	225.385756	574.402505
	Hombre	43769.0	**121.706488**	46.466150	75.071609	101.188915	**113.388524**	128.470969	596.345931
Sesenta	**Mujer**	21219.0	**176.181293**	25.591177	83.378015	163.296771	**171.553172**	181.499779	596.375817
	Hombre	31708.0	**132.557706**	65.464476	75.546207	100.920384	**117.020848**	136.392822	595.687572

Tabla 5.6: Descripción de la mediana de la frecuencia fundamental en grupos de edades en el idioma chino.

Edad	Género	Cantidad	Media	Desv. est.	Mínimo	Primer cuartil	Mediana	Tercer cuartil	Máximo
Veinte	**Mujeres**	5840.0	**222.474257**	35.744485	104.013892	203.925776	**222.398512**	240.623530	540.994143
	Hombre	23070.0	**136.917224**	39.501145	80.617342	118.064250	**130.020726**	144.962066	583.096654
Treinta	**Mujeres**	2581.0	**215.496858**	25.201018	116.996809	203.594217	**216.184941**	226.792505	386.994340
	Hombre	8341.0	**148.987706**	45.772433	78.880281	117.655217	**135.591320**	173.175065	562.267153
Cuarenta	**Mujeres**	2956.0	**216.932013**	33.926677	129.878111	193.720965	**210.502681**	235.508688	538.666197
	Hombre	3405.0	**130.668720**	38.058882	76.408887	108.729769	**131.676341**	149.171463	575.064418
Cincuenta	**Mujeres**	14.0	**213.196560**	18.289266	188.648687	198.997376	**209.968166**	226.404648	242.761580
	Hombre	135.0	**146.883945**	54.376210	80.400703	115.243991	**134.014866**	157.696631	420.701886
Sesenta	**Hombre**	7.0	**195.469211**	129.764640	114.996347	126.563879	**128.990198**	200.335441	470.499297

Tabla 5.7: Descripción de la mediana de la frecuencia fundamental en grupos de edades en los idiomas español, alemán, francés, inglés y chino.

Edad	Género	Cantidad	Media	Desv. est.	Mínimo	Primer cuartil	Mediana	Tercer cuartil	Máximo
Veinte	**Mujer**	122026.0	**219.168161**	33.592543	79.348518	198.736992	**218.404380**	238.397388	571.533359
	Hombre	491196.0	**131.593098**	40.380337	75.142126	111.649714	**124.333444**	140.813233	597.071396
Treinta	**Mujer**	51105.0	**201.914810**	31.866760	82.654962	183.637206	**199.969781**	217.863730	572.116091
	Hombre	372163.0	**126.968408**	40.901994	75.292111	108.986046	**120.626837**	133.812643	597.414785
Cuarenta	**Mujer**	41923.0	**203.096796**	38.904783	82.067515	179.124076	**200.022459**	227.108340	595.367925
	Hombre	321550.0	**120.459674**	43.017037	75.440183	97.064241	**113.208511**	131.592027	598.393021
Cincuenta	**Mujer**	45438.0	**198.344715**	38.701051	84.078296	172.318588	**196.567541**	217.923472	574.402505
	Hombre	186211.0	**116.074173**	35.754713	75.071609	98.823466	**108.672596**	123.862852	596.345931
Sesenta	**Mujer**	30047.0	**184.427493**	29.742117	78.051816	165.939233	**175.816061**	198.977295	596.375817
	Hombre	122585.0	**134.975049**	47.780745	75.546207	118.667187	**128.989593**	140.128232	595.687572

- **Resultados**

  Los resultados de la clasificación de género de la primera metodología pueden verse en la Tabla 5.8. La discusión por escenario es la siguiente:

  1. Escenario de voces repetidas y no balanceadas (denotada como **NB-Repetidas**): El clasificador que mejor se desempeñó en todos los idiomas fue el J48. Teniendo el puntaje F1 más alto (98.9 %) en el idioma Español y su puntaje F1 más bajo (97.2 %) en los idiomas Inglés y Chino. Para el caso de todos los idiomas, obtuvo un puntaje F1 de 95.5 %.
  2. Escenario de voces repetidas y balanceadas (denotada como **B-Repetidas**): El algoritmo J48 obtuvo la precisión más alta (96.8 %) en el idioma Español y la más baja en el idioma Chino (94.2 %). Para el caso de todos los idiomas, el perceptrón multicapa obtuvo 94.2 % de precisión.
  3. Escenario de voces únicas y no balanceadas (denotada como **NB-Únicas**): El algoritmo perceptrón multicapa obtuvo el puntaje F1 más alto (98.5 %) en el idioma Alemán y la el más bajo en el idioma Español (95.7 %). Para el caso de todos los idiomas, el perceptrón multicapa obtuvo 97.2 % de puntaje F1.
  4. Escenario de voces únicas y balanceadas (denotada como **B-Únicas**): El algoritmo MLP obtuvo la mayor precisión (95.5 %) en el idioma Francés y la más baja en el idioma Español (94.0 %).

Tabla 5.8: Resultados de la clasificación de género en varios idiomas 75-600 Hertz.

Lenguaje	Algoritmo	NB-Repetidas			B-repetidas			NB-Únicas			B-Únicas		
		Puntaje F1	**Exhaustividad**	**Área ROC**	**Precisión**	**Exhaustividad**	**Área ROC**	**Puntaje F1**	**Exhaustividad**	**Área ROC**	**Precisión**	**Exhaustividad**	**Área ROC**
Español	**LR**	97.2 %	98.5	96.1 %	90.9 %	92.2 %	95.4 %	95.3 %	97 %	95.2 %	**94.0 %**	94.8 %	97.5 %
	MLP	98.7 %	98.8 %	98.8 %	96 %	96.6 %	98.8 %	**95.7 %**	95.9 %	96.2 %	91.5 %	91.7 %	96 %
	J48	**98.8 %**	98.8 %	98.4 %	**96.8 %**	96.1 %	98.4 %	95.3 %	96 %	91.3 %	91.4 %	92.35	92.6 %
Inglés	**LR**	96.2 %	97.0 %	97.4 %	93.9 %	93.5 %	97.6 %	95.1	96.7 %	95.7 %	92.4	92.5 %	95.8 %
	MLP	96.9 %	96.8 %	98.1 %	**95.1 %**	95.1 %	98.2 %	**96.4 %**	96 %	96.7 %	93.4 %	92.8 %	96.4 %
	J48	**97.2 %**	97.2 %	97.8 %	**92.2**	93.5 %	94.4	**96.4 %**	96.2 %	93.7 %	**94.9**	94.9	96.9
Alemán	**LR**	98.1 %	99 %	98.1 %	95.6	95.4 %	98.4 %	97.9 %	98.8 %	98.1 %	**95.4**	95.4 %	98.6 %
	MLP	**98.9 %**	98.8 %	98.8 %	95.8 %	96.7 %	98.9 %	98.3 %	98.3 %	98.2 %	94.3 %	94.9 %	97.9
	J48	**98.9 %**	98.9 %	98.1 %	**96.4 %**	96.3 %	98.1 %	**98.5 %**	98.8 %	92.6 %	94.2 %	93.8 %	95.1 %
Francés	**LR**	98.2 %	99 %	97.9 %	93.6 %	95.1 %	98.2 %	97.1 %	98.1 %	97.2 %	95.4	95.8 %	97.5 %
	MLP	**98.7 %**	98.9 %	98.6 %	**96 %**	94.3 %	95.16 %	97.1 %	98.1 %	97.2 %	**95.5 %**	94.2 %	96.6 %
	J48	**98.7 %**	99 %	97.4 %	95.3 %	95.2	97.7 %	**97.8 %**	98.1 %	93.9 %	94.1 %	96 %	95.2 %
Chino	**LR**	96.1 %	96.6 %	96.3 %	**93.9 %**	93.2 %	96.8 %	95.4 %	96.9 %	95.8 %	**94.4 %**	92.3 %	96.4 %
	MLP	97 %	97 %	97.3 %	93.5 %	95.2 %	97.4 %	**96.2 %**	96.2 %	97 %	92.2 %	90.7 %	96.3 %
	J48	**97.2 %**	97.3 %	96.8 %	93.9 %	95 %	96.9 %	95.9 %	95.7 %	90.9 %	93.9 %	92.9 %	93.8 %
Todos	**LR**	94.5 %	94.6 %	97.2 %	92.5 %	92.8 %	96.3 %	95.7 %	97.3 %	96.1 %	93.4 %	94.1 %	97.4 %
	MLP	93.5 %	93.4 %	96.4 %	**94.2 %**	95.1 %	98.0 %	**97.2 %**	97.2 %	97.3 %	93.2 %	93.2 %	97 %
	J48	**95.5 %**	95.5 %	97.7 %	**92.8 %**	93.2 %	94.9 %	96.7 %	96.5 %	94.5 %	**94.6 %**	94.9 %	97.8 %

5.2. Segundo Caso de Prueba del Reconocimento de Género por Voz

- **Introducción**
  La presencia de ruido en los ambientes no controlados, modifica las estimaciones de F_0 de un hablante. Por lo que se requieren de filtros para eliminar los audios que no reflejen los valores o rango de valores esperados de F_0. La segunda metodología denominada atenuación del error de octava (ver Figura 5.2) es una variación de la primera en donde se realiza una atenuación del error de octava cometido. Esto es, de los audios calculados, se descartaban aquellos cuyo valor de la mediana fuera mayor a 350 Hertz[1].
- **Objetivo** El objetivo de la metodología atenuación del ruido, es analizar el desempeño de las características estadísticas derivadas de la frecuencia fundamental y atenuadas al error de octava, en el reconocimiento de género.
  Los objetivos específicos son:
  - Realizar el reconocimiento de género en los idiomas Inglés, Español, Alemán, Francés, Chino y su combinación por medio de las características estadísticas mínimo de F_0, primer cuartil de F_0, media de F_0, mediana de F_0, tercer cuartil de F_0, Máximo de F_0, desviación estándar de F_0, Rango intercuartil de F_0 y la estimación de la edad en décadas.
  - Validar las estadísticas empleadas mediante el estudio de escenarios balanceados y no balanceados donde cada uno tendrá los casos de voces repetidas y voces únicas.
  - Emplear tres tipos diferentes de clasificadores (Perceptrón multicapa, Árbol de búsqueda J48, y regresión lógística) para comparar las métricas obtenidas.

[1] Esta idea fue sugerída por el Dr. Santiago Barrera.

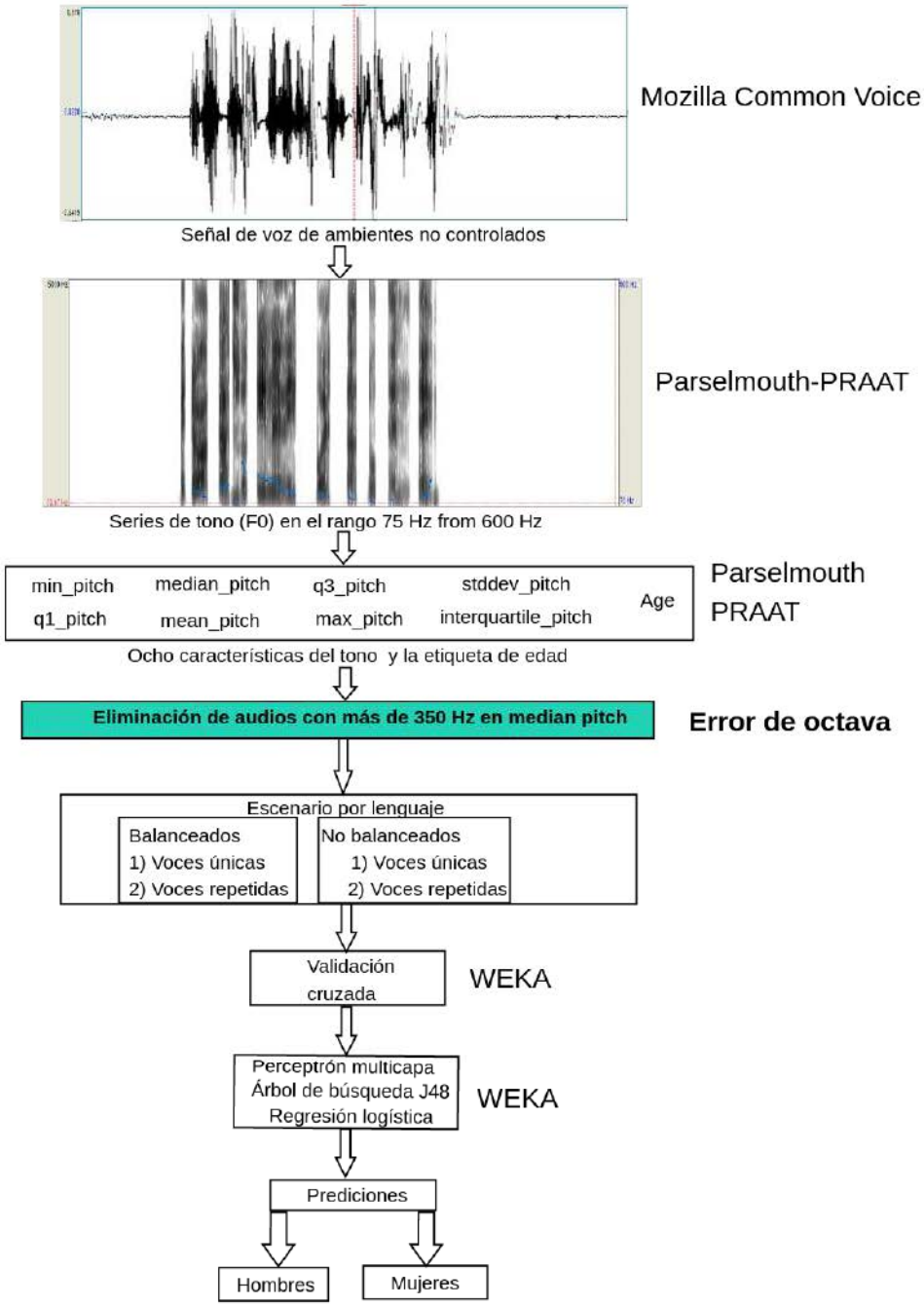


Figura 5.2: Metodología 2 propuesta para el reconocimiento de género por voz mediante algoritmos de aprendizaje clásicos.

- **Características de las pruebas realizadas**
  Los conjuntos de datos fueron los mismos que en la primer metodología.
- **Resultados**
  Los resultados de la clasificación de género de la segunda metodología pueden verse en la Tabla 5.9. La discusión por escenario es la siguiente:
  1. Escenario de voces repetidas y no balanceadas (**NB-Repetidas**): Para este escenario tanto el perceptrón multicapa como el J48 obtuvieron puntajes F1 similares. El idioma que mejor clasificaron ambos fue el Alemán con 97.6 % de puntaje F1 y el de peor puntaje fue el Chino e Inglés con 95.5 %. En el caso de todos los idiomas, el J48 obtuvo el puntaje más alto con 96.2 %.
  2. Escenario de voces repetidas y balanceadas (**B-Repetidas**): El algoritmo perceptrón multicapa obtuvo la precisión más alta (96.4 %) en el idioma Español y la más baja en el idioma Inglés (93.7 %). Para el caso de todos los idiomas, tanto el perceptrón multicapa como J48 obtuvo un 94.1 % de precisión.
  3. Escenario de voces únicas y no balanceadas (**NB-Únicas**): El algoritmo J48 obtuvo el puntaje F1 más alto (97.0 %) en el idioma Alemán y el más bajo en el idioma Español (93 %). Para el caso de todos los idiomas, el perceptrón multicapa obtuvo un 94.6 % de puntaje F1.
  4. Escenario de voces únicas y balanceadas (**B-Únicas**): El algoritmo regresión logística obtuvo la mayor precisión (96.2 %) en el idioma Francés y la más baja en el idioma Español (92.1 %). Para el caso de todos los idiomas, el perceptrón multicapa obtuvo un 92.8 % de precisión.

Lenguaje	Algoritmo	NB-Repetidas			B-Repetidas			NB-Únicas			B-Únicas		
		Puntaje F1	**Exhaustividad**	**Área ROC**	**Precisión**	**Exhaustividad**	**Área ROC**	**Puntaje F1**	**Exhaustividad**	**Área ROC**	**Precisión**	**Exhaustividad**	**Área ROC**
Español	**LR**	97.1 %	97.0 %	98.7 %	95.8 %	95.8 %	98.7 %	**93 %**	93 %	96.8 %	**92.1 %**	92 %	96.9 %
	MLP	**97.6 %**	97.6 %	98.9 %	**96.4 %**	96.4 %	98.9 %	91.5 %	91.5 %	94.9 %	91.4 %	91.4 %	95.9 %
	J48	**97.6 %**	97.6 %	97.2 %	96.3 %	96.3 %	98 %	92.5 %	92.5 %	90.7 %	91.3 %	91.3 %	91.1 %
Inglés	**LR**	94.3 %	94.3 %	97.6 %	93.2 %	93.2 %	97.6 %	93.7 %	93.8 %	96.7 %	92.5 %	92.5 %	97.0 %
	MLP	95 %	95 %	98.0 %	**93.7 %**	93.7 %	97.9 %	**93.9 %**	93.9 %	96.5 %	93.8 %	93.8 %	97 %
	J48	**95.5 %**	95.5 %	97.2 %	93.6 %	93.6 %	97.1 %	93.8 %	93.7 %	92.8 %	**94.2 %**	94.2 %	96.6 %
Alemán	**LR**	97.4 %	97.4 %	98.5 %	95.8 %	95.8 %	98.7 %	**96.9 %**	96.9 %	98.3 %	**94.9 %**	94.9 %	98.1 %
	MLP	**97.7 %**	97.7 %	98.8 %	96.1 %	96.1 %	98.9 %	**96.3 %**	96.3 %	98.0 %	94.1 %	94.1 %	97.2 %
	J48	**97.7 %**	97.7 %	97.8 %	**96.3 %**	96.3 %	97.8 %	**97.0 %**	97.0 %	92.0 %	94.1 %	94.1 %	95 %
Francés	**LR**	97.3 %	97.4 %	97.4 %	94.7 %	94.6 %	98.6 %	**96.5 %**	96.5 %	98.6 %	**96.2 %**	96.2 %	98.8 %
	MLP	**97.5 %**	97.6 %	97.6 %	**95.3 %**	95.3 %	98.8 %	96.0 %	96.0 %	98.2 %	95.4 %	95.4 %	98.5 %
	J48	**97.6 %**	97.6 %	97.2 %	95.0 %	95.0 %	97.9 %	95.7 %	95.7 %	93.8 %	95.8 %	95.8 %	94.6 %
Chino	**LR**	94.6 %	94.6 %	97.1 %	93.3 %	93.3 %	96.9 %	92.1 %	92.1 %	96.4 %	89.6 %	89.6	96.2 %
	MLP	**95.5 %**	95.5 %	97.5 %	**94.2 %**	94.2 %	97.5 %	**93.1 %**	93.2 %	97 %	**91.5 %**	91.5 %	94.9 %
	J48	**95.5 %**	95.6 %	96.6 %	94 %	94 %	95.5 %	91.6 %	91.6 %	90.8 %	90.9 %	90.9 %	87.6 %
Todos	**LR**	95.6 %	95.6 %	97.6 %	93.5 %	93.5 %	97.6 %	94.3 %	94.4 %	97.2 %	92.5 %	92.5 %	92.5 %
	MLP	96.1 %	96.1 %	97.8 %	**94.1 %**	94.1 %	98.0 %	**94.6 %**	94.6 %	97.0 %	**92.8 %**	92.8 %	96.8 %
	J48	**96.2 %**	96.2 %	97.5 %	**94.1 %**	94.1 %	96.9 %	94.2 %	94.2 %	94.0 %	92.6 %	92.6 %	94.2 %

Tabla 5.9: Métricas de los clasificadores usados en el reconocimiento de género en varios idiomas 75-350 Hertz.

5.3. Tercer Caso de Prueba del Reconocimento de Género por Voz

- **Introducción**

  Uno de los aspectos esenciales de los sistemas de reconocimiento de género por voz es utilizar características que preserven la distinción de géneros independientemente del idioma en que fueron entrenados y del tipo de ambiente de grabación (controlado o no controlado). Las estadísticas mostradas en la primera metodología señalan que el idioma efectivamente repercute en el rango de valores de la frecuencia fundamental, pero es posible combinar la información de F_0 con otras características biológicas del hablante. La combinación de F_0 con la estimación de la longitud del tracto vocal y los formantes pueden ser una alternativa, debido a que brindan un perfil más distintivo de un género. Otra alternativa, son los coeficientes cepstrales de frecuencias de Mel por su uso en el reconocimiento del habla, ya que puede inferir información acústica del hablante.

- **Objetivo**

  El objetivo general de esta metodología fue comparar el desempeño en el reconocimiento de género por voz de los tres grupos (tono, tono y tracto vocal, y grupos de $MFCC$) de características analizadas en el presente trabajo en tres distintos conjuntos de voces donde dos de ellos son en inglés y uno es en español.

  Como objetivos específicos se tienen:

  1. Detección de género mediante cada grupo de características tonales, formantes y cepstrales, en cada uno de los diferentes conjuntos de datos (Español, Inglés Mozilla e Inglés $PVQD$).
  2. Validación de los modelos generados en un mismo grupo de descriptores con los dos conjuntos de datos restantes semejantes (i.e. mismas características).

  De esta forma se obtienen 5 vectores de características. Esta división de los últimos 3 grupos de características, se debe a que han sido utilizados en el reconocimiento de género por voz ([3, 82, 83]). Consecuentemente, mediante una validación cruzada de 10 carpetas se entrena al algoritmo clasificador para obtener las predicciones hombre o mujer.

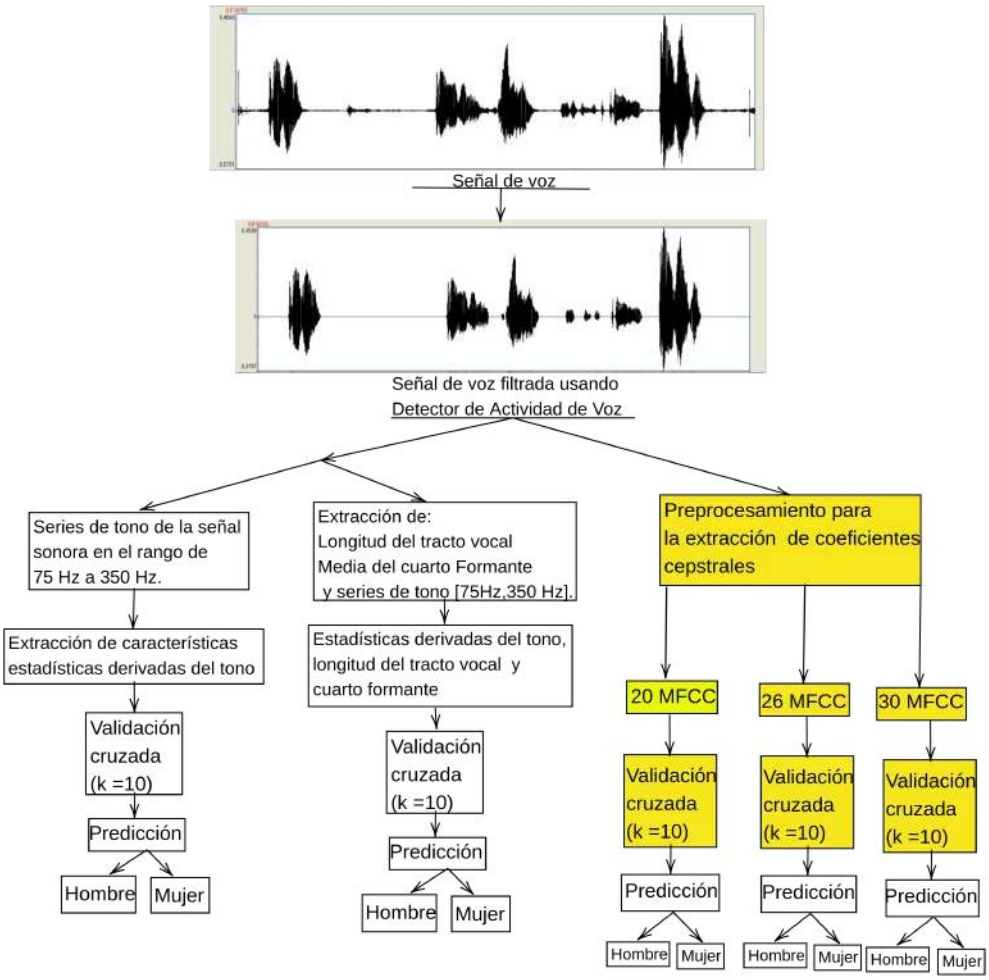


Figura 5.3: Metodología para el reconocimiento de género por voz usando características tonales, formantes y cepstrales.

Una vez obtenido el modelo en un grupo de características, se valida dicho modelo mediante la evaluación de su desempeño en un segundo idioma y ambiente (véase la Fig. 5.4).

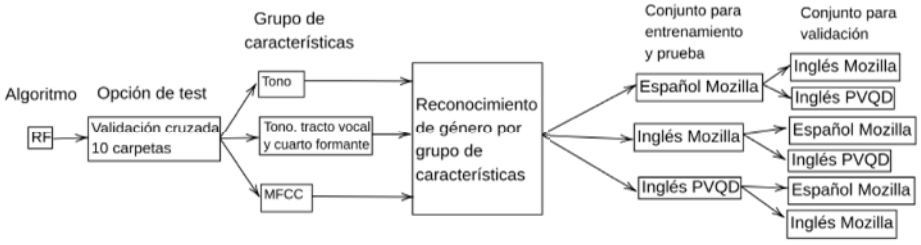


Figura 5.4: Validación de los modelos en diferentes conjuntos de datos.

- **Características de las Pruebas Realizadas**
  Los conjuntos de voces provienen de Mozilla [13] en los idiomas Español (182, 574 audios con una duración de 3.29 seg. y una desviación estándar de 1.13 seg.), Inglés (73, 278 audios con una duración de 2.164 seg. y una desviación estándar de 1.083 seg.), y la base de datos de cualidades perceptivas de la voz o *Perceptual Voice Quality Dataset PVQD* [75] (296 audios con una duración de 24.48 seg.

y una desviación estándar de 7.609 seg.). Una descripción estadística de cada conjunto puede verse en las Tablas 5.10, 5.11 y 5.12. A cada una de las características presentes en los 5 grupos se realizó un proceso de *Winsorización* [72]. Este proceso consiste en sustituir los datos atípicos por una media *winsorizada*. La media winsorizada es una medida de tendencia central robusta que se utiliza para resumir un conjunto de datos que puede contener valores atípicos (*outliers*) o valores extremos. La media winsorizada es una versión modificada de la media aritmética que se calcula después de reemplazar los valores extremos por los valores más cercanos dentro de un rango predefinido. Para calcular la media winsorizada, se siguen los siguientes pasos:

- Se ordena el conjunto de datos de menor a mayor.
- Se establece un porcentaje de recorte (por ejemplo, el 20 %).
- Se reemplazan los valores extremos en ambos extremos del conjunto de datos por los valores más cercanos dentro del rango de recorte.
- Se calcula la media aritmética de los datos modificados.

Posteriormente, cada dato fuera del porcentaje de recorte es sustituido por dicha media. El porcentaje de recorte para cada característica fue de 5 %. De este modo se garantiza que el 95 % de las observaciones sean representativas mientras que el resto, son datos provenientes de la media *winsorizada*.

Tabla 5.10: Distribución de la mediana del tono y la estimación del tracto vocal (ambos *winsorizados*) en el conjunto de voces en Español (entrenamiento).

	Mediana del tono						
Género	**Cantidad**	**Mínimo**	**Primer cuartil**	**Mediana**	**Media**	**Tercer cuartil**	**Máximo**
Hombre	130113	77.03	**110.88**	**124.19**	124.36	135.75	336.55
Mujer	52461	77.97	**189.46**	**202.45**	205.20	221.25	332.61
	Estimación del tracto vocal						
Hombre	130113	8.93	**16.12**	**17.09**	17.07	18.06	26.77
Mujer	52461	9.14	**15.15**	**16.22**	16.99	19.49	22.82

Tabla 5.11: Distribución de la mediana del tono y la estimación del tracto vocal (ambos *winsorizados*) en el conjunto de voces en Inglés (entrenamiento).

	Mediana del tono						
Género	**Cantidad**	**Mínimo**	**Primer cuartil**	**Mediana**	**Media**	**Tercer cuartil**	**Máximo**
Hombre	55029	75.04	**104.99**	**117.66**	122.25	134.40	322.44
Mujer	18249	78.25	**178.27**	**198.90**	199.97	220.98	338.45
	Estimación del tracto vocal						
Hombre	55029	8.78	**16.87**	**17.65**	17.30	18.39	26.27
Mujer	18249	8.56	**14.80**	**15.41**	15.41	16.07	26.53

- **Resultados** A continuación, se mostrarán los resultados en dos etapas: La evaluación del reconocimiento por idioma y la validación de los modelos en un segundo idioma y ambiente.

Tabla 5.12: Distribución de la mediana del tono y la estimación del tracto vocal (ambos *winsorizados*) en el conjunto de voces en Inglés en ambientes controlados.

	Mediana del tono						
Género	**Cantidad**	**Mínimo**	**Primer cuartil**	**Mediana**	**Media**	**Tercer cuartil**	**Máximo**
Hombre	100	79.62	**103.08**	**119.12**	132.29	148.24	286.64
Mujer	196	96.76	**173.12**	**196.26**	195.11	214.74	272.97
	Estimación del tracto vocal						
Hombre	100	10.16	**16.26**	**17.11**	16.86	17.83	19.94
Mujer	196	9.74	**14.20**	**14.68**	14.47	15.08	17.63

5.3.1. Evaluación de Reconocimiento por Idioma

En la Tabla 5.13, se muestran los resultados por escenarios, género e idioma. Para identificar cada modelo creado, se utilizó la siguiente notación

$$Algoritmo\_Car\_Train\_Idioma,$$

donde la primera parte denomina el algoritmo utilizado (RF en este caso), la segunda parte Car, es el grupo de características (Tono, Tono + VTL[2], MFCC20, MFCC26 y MFCC30) y el Idioma (Español como Esp, Inglés Mozilla como Ing, e Inglés $PVQD$ como IngC). Los puntos más revelantes de la Tabla 5.13 se exponen a continuación:

1. Las características Tono + VTL obtuvieron el mejor desempeño en las métricas de precisión (0.989 en Español, 0.973 en Inglés Mozilla y 0.924 en Inglés $PVQC$) y puntaje F1 (0.99 en Español, 0.973 en Inglés Mozilla y 0.885 en Inglés $PVQC$) en la categoría de los hombres.
2. El grupo de 30 $MFCC$ obtuvieron el mejor desempeño en precisión (0.975 en Español, 0.927 en Inglés Mozilla y 0.851 en Inglés $PVQC$) y Puntaje F1 (0.984 en Español, 0.925 en Inglés Mozilla y 0.791 en Inglés $PVQC$) en la categoría de hombres con respecto al grupo de 26 y 20.
3. Los resultados del punto 1 y 2 son semejantes en la categoría de las mujeres. Esto es, las características Tono + VTL y $MFCC$ 30 obtuvieron los mejores desempeños en precisión y puntaje F1 en la categoría de mujeres.
4. El escenario con mayor exactitud fue Tono + VTL (0.986 en Español, 0.959 en Inglés Mozilla y 0.926 en Inglés $PVQC$).

[2] VTL son las siglas de *Vocal Tract Length*. Este escenario es la combinación de caraterísticas tonales, la media del cuarto formante y la estimación del tracto vocal.

Tabla 5.13: Distribución de las métricas en los tres grupos de características y los tres conjuntos de voces (Español-Esp, Inglés Mozilla-Ing, e Inglés *PVQD*-IngC).

Escenario	Precisión hombres	Exhaustividad hombres	Puntaje F1 Hombres	Precisión mujeres	Exhaustividad mujeres	Puntaje F1 mujeres	Exactitud	Área bajo la curva ROC hombres	Área bajo la curva ROC mujeres
RF_Tono_Train_Esp	0.982	0.987	0.985	0.967	0.956	0.961	0.978	0.994	0.994
RF_Tono +VTL_Train_Esp	**0.989**	**0.991**	**0.99**	**0.977**	**0.973**	**0.975**	**0.986**	**0.997**	**0.997**
RF_MFCC20_Train_Esp	0.968	0.987	0.977	0.965	0.919	0.941	0.967	0.995	0.995
RF_MFCC26_Train_Esp	0.972	0.99	0.981	0.973	0.929	0.951	0.972	0.996	0.996
RF_MFCC30_Train_Esp	**0.975**	**0.992**	**0.984**	**0.98**	**0.938**	**0.958**	**0.977**	**0.997**	**0.997**

Escenario	Precisión hombres	Exhaustividad hombres	Puntaje F1 Hombres	Precisión mujeres	Exhaustividad mujeres	Puntaje F1 mujeres	Exactitud	Área bajo la curva ROC hombres	Área bajo la curva ROC mujeres
RF_Tono_Train_Ing	0.959	0.955	0.957	0.866	0.877	0.872	0.936	0.972	0.972
RF_Tono +VTL_Train_Ing	**0.973**	**0.972**	**0.973**	**0.916**	**0.919**	**0.917**	**0.959**	**0.989**	**0.989**
RF_MFCC20_Train_Ing	0.918	0.977	0.946	0.912	0.736	0.814	0.916	0.967	0.967
RF_MFCC26_Train_Ing	0.925	0.980	0.952	0.926	0.760	0.835	0.925	0.975	0.975
RF_MFCC30_Train_Ing	**0.927**	**0.927**	**0.925**	**0.927**	**0.769**	**0.840**	**0.928**	**0.976**	**0.976**

Escenarios	Precisión hombres	Exhaustividad hombres	Puntaje F1 Hombres	Precisión mujeres	Exhaustividad mujeres	Puntaje F1 mujeres	Exactitud	Área bajo la curva ROC hombres	Área bajo la curva ROC mujeres
RF_Tono_Train_IngC	0.816	0.71	0.759	0.861	0.918	0.889	0.849	0.882	0.882
RF_Tono +VTL_Train_IngC	**0.924**	**0.85**	**0.885**	**0.926**	**0.964**	**0.945**	**0.926**	**0.943**	**0.943**
RF_MFCC20_Train_IngC	0.805	0.66	0.725	0.841	0.918	0.878	0.831	0.881	0.881
RF_MFCC26_Train_IngC	0.814	0.7	0.753	0.857	0.918	0.887	0.845	0.895	0.895
RF_MFCC30_Train_IngC	**0.851**	**0.74**	**0.791**	**0.876**	**0.934**	**0.904**	**0.868**	**0.909**	**0.909**

5.3.2. Validación de los Modelos

Los resultados de la validación de los modelos pueden verse en las Tablas 5.14, 5.15, 5.16, 5.17, 5.18 y 5.19. Sin embargo, se hará una discusión de cada validación en las siguientes subseciones.

5.3.3. Entrenamiento en Español y Validación en Inglés Mozilla e Inglés PVQD

Los siguientes puntos detallan los aspectos más relevantes de la tabla 5.14.

1. Las características Tono + *VTL* obtuvieron el mayor desempeño de Puntaje F1 tanto en hombres (0.964 en Inglés Mozilla y 0.806 en Inglés *PVQC*) como en mujeres (0.892 en Inglés Mozilla y 0.896 en Inglés *PVQC*).
2. El grupo de los coeficientes cepstrales obtuvieron un desempeño de puntaje F1 en hombres (0.893 en Inglés Mozilla y 0.643 en Inglés *PVQC*) y en mujeres (0.638 en Inglés Mozilla y 0.673 en Inglés *PVQC*), mayor al 63 % pero menor al 90 %. Esto indica su sensibilidad al cambio de idioma.
3. En cuanto al cambio de ambientes no controlados a controlados, si se compara el puntaje F1 de hombres y mujeres de *RF* _Tono+*VTL*_*Train*_IngC de la Tabla 5.13 con el de *RF* _Tono+*VTL*_*Train*_Esp_valid_IngC de la tabla 5.14, se observa una diferencia de 0.079 para hombres y 0.049 para mujeres. Eso significa que hay un cambio de aproximadamente del 8 % y del 4 % en el reconocimiento de hombres y mujeres, respectivamente. Mientras que si se realiza la misma comparación con *MFCC* 30, se obtiene una diferencia del 15 % y del 23 %, respectivamente (véase la Tabla 5.15). Por lo anterior, se muestra que el cambio de ambiente repercute en el desempeño del modelo.

Tabla 5.14: Distribución de las métricas del entrenamiento en Español Mozilla y validación en Inglés Mozilla e Inglés PVQD.

Algoritmos	Precisión hombres	Exhaustividad hombres	Puntaje F1 Hombres	Precisión mujeres	Exhaustividad mujeres	Puntaje F1 mujeres	Exactitud	Área bajo la curva ROC hombres	Área bajo la curva ROC mujeres
RF_Tono_Train_Esp_valid_Ing	0.929	0.96	0.944	0.873	0.786	0.827	0.9156	0.965	0.965
RF_Tono +VTL_Train_Esp_valid_Ing	**0.954**	**0.974**	**0.964**	**0.921**	**0.865**	**0.892**	**0.9463**	**0.98**	**0.98**
RF_MFCC20_Train_Esp_valid_Ing	0.853	0.896	0.874	0.648	0.554	0.597	0.8083	0.831	0.831
RF_MFCC26_Train_Esp_valid_Ing	0.855	0.909	0.882	0.678	0.554	0.61	0.8181	0.852	0.852
RF_MFCC30_Train_Esp_valid_Ing	**0.862**	**0.926**	**0.893**	**0.726**	**0.569**	**0.638**	**0.8345**	**0.871**	**0.871**
Algoritmos	**Precisión hombres**	**Exhaustividad hombres**	**Puntaje F1 Hombres**	**Precisión mujeres**	**Exhaustividad mujeres**	**Puntaje F1 mujeres**	**Exactitud**	**Área bajo la curva ROC hombres**	**Área bajo la curva ROC mujeres**
RF_Tono_Train_Esp_valid_IngC	0.686	0.83	0.751	0.903	0.806	0.852	0.814	0.874	0.874
RF_Tono +VTL_Train_Esp_valid_IngC	**0.783**	**0.83**	**0.806**	**0.911**	**0.883**	**0.896**	**0.865**	**0.921**	**0.921**
RF_MFCC20_Train_Esp_valid_IngC	0.428	0.92	0.584	0.901	0.372	0.527	0.557	0.809	0.809
RF_MFCC26_Train_Esp_valid_IngC	0.456	0.93	0.612	0.924	0.434	0.59	0.601	0.852	0.852
RF_MFCC30_Train_Esp_valid_IngC	**0.497**	**0.91**	**0.643**	**0.92**	**0.531**	**0.673**	**0.659**	**0.875**	**0.875**

Tabla 5.15: Diferencias entre el aprendizaje en ambiente controlado y el aprendizaje en el ambiente no controlado en el idioma Español.

	RF_Tono+VTL_ Train_IngC	RF_Tono+VTL_ Train_Esp_valid_IngC	Diferencia	Aproximación en porcentaje
Puntaje F1 Hombres	**0.885**	**0.806**	**0.079**	**8 %**
Puntaje F1 Mujeres	**0.945**	**0.896**	**0.049**	**5 %**
Exactitud	**0.926**	**0.865**	**0.061**	**6 %**
	RF_MFCC30_ Train_IngC	**RF_MFCC30_Train_Esp_valid_IngC**	**Diferencia**	**Aproximación en porcentaje**
Puntaje F1 Hombres	**0.791**	**0.643**	**0.148**	**15 %**
Puntaje F1 Mujeres	**0.904**	**0.673**	**0.231**	**23 %**
Exactitud	**0.868**	**0.659**	**0.209**	**21 %**

5.3.4. Entrenamiento en Inglés Mozilla y validación en Español e Inglés PVQC

Los siguientes puntos detallan los aspectos más relevantes de la tabla 5.16.

1. Las características Tono + *VTL* obtuvieron el mayor desempeño de Puntaje F1 tanto en hombres (0.971 en Español Mozilla y 0.879 en Inglés *PVQC*) como en mujeres (0.902 en Español Mozilla y 0.715 en Inglés *PVQC*).
2. El grupo de los 30 coeficientes cepstrales obtuvieron un desempeño de puntaje F1 tanto en hombres (0.926 en Español, 0.624 en Inglés *PVQD*) como en mujeres (0.715 en Español y 0.619 en Inglés *PVQD*), mayor al 61 % pero menor al 93 %. Nuevamente, se muestra otra evidencia de su sensibilidad al cambio de idioma.
3. En cuanto al cambio de ambientes no controlados a controlados (véase la Tabla 5.19), la diferencia de puntaje F1, aproximada en porcentaje, tanto en hombres como en mujeres de los modelos RF_Tono+*VTL_Train_* IngC y RF_Tono+*VTL_Train_* Esp_ valid_IngC es de 8 % y 5 %, respectivamente. Para el uso de los 30 *MFCC* las diferencias de puntajes F1 aproximadas fueron -15 %[3] en hombres y 28 % en mujeres. Por lo anterior, se tiene otra evidencia de que el cambio del ambiente repercutió en el desempeño del clasificador.

[3]Esto significa que aprendió mejor los ejemplos del Inglés *PVQD* mediante el entrenamiento del Inglés de Mozilla. Para este caso, las métricas del RF_ MFCC30_ Train_ IngC son consideradas como el valor estándar, y la fórmula para calcular el porcentaje es $(estándar - valor/estándar)*100$.

Tabla 5.16: Distribución de las métricas del entrenamiento en Inglés Mozilla y validación en Español e Inglés PVQD.

Algoritmos	Precisión hombres	Exhaustividad hombres	Puntaje F1 Hombres	Precisión mujeres	Exhaustividad mujeres	Puntaje F1 mujeres	Exactitud	Área bajo la curva ROC hombres	Área bajo la curva ROC mujeres
RF_Tono_Train_Ing_valid_Esp	0.972	0.949	0.96	0.836	0.907	0.87	0.939	0.974	0.974
RF_Tono +VTL_Train_Ing_valid_Esp	**0.978**	**0.964**	**0.971**	**0.88**	**0.926**	**0.902**	**0.955**	**0.984**	**0.984**
RF_MFCC20_Train_Ing_valid_Esp	0.868	0.969	0.916	0.82	0.49	0.613	0.862	0.885	0.885
RF_MFCC26_Train_Ing_valid_Esp	0.88	0.962	0.919	0.804	0.545	0.649	0.868	0.910	0.910
RF_MFCC30_Train_Ing_valid_Esp	**0.906**	**0.947**	**0.926**	**0.781**	**0.659**	**0.715**	**0.882**	**0.924**	**0.924**

Algoritmos	Precisión hombres	Exhaustividad hombres	Puntaje F1 Hombres	Precisión mujeres	Exhaustividad mujeres	Puntaje F1 mujeres	Exactitud	Área bajo la curva ROC hombres	Área bajo la curva ROC mujeres
RF_Tono_Train_Ing_valid_IngC	0.737	0.84	0.785	0.912	0.847	0.878	0.845	0.889	0.889
RF_Tono +VTL_Train_Ing_valid_IngC	**0.888**	**0.87**	**0.879**	**0.934**	**0.944**	**0.939**	**0.919**	**0.931**	**0.930**
RF_MFCC20_Train_Ing_valid_IngC	0.384	0.99	0.553	0.974	0.189	0.316	0.459	0.766	0.766
RF_MFCC26_Train_Ing_valid_IngC	0.413	0.97	0.579	0.951	0.296	0.451	0.524	0.841	0.841
RF_MFCC30_Train_Ing_valid_IngC	**0.47**	**0.93**	**0.624**	**0.929**	**0.464**	**0.619**	**0.622**	**0.851**	**0.851**

Tabla 5.17: Diferencias entre el aprendizaje en ambiente controlado y el aprendizaje en el ambiente no controlado en el idioma Inglés.

	RF_Tono+VTL_Train_IngC	RF_Tono+VTL_Train_Ing_valid_IngC	Diferencia	Aproximación en porcentaje
Puntaje F1 Hombres	**0.885**	**0.879**	**0.006**	**0.6 %**
Puntaje F1 Mujeres	**0.945**	**0.624**	**0.321**	**32 %**
Exactitud	**0.926**	**0.919**	**0.007**	**0.01 %**
	RF_MFCC30_Train_IngC	**RF_MFCC30_Train_Ing_valid_IngC**	**Diferencia**	**Aproximación en porcentaje**
Puntaje F1 Hombres	**0.791**	**0.939**	**-0.148**	**-15 %* Mejora**
Puntaje F1 Mujeres	**0.904**	**0.619**	**0.285**	**28 %**
Exactitud	**0.868**	**0.622**	**0.246**	**25 %**

5.3.5. Entrenamiento en Inglés PVQD y validación en Español e Inglés Mozilla

Los siguientes puntos detallan los aspectos más relevantes de la tabla 5.18.

1. Las características Tono + *VTL* obtuvieron el mayor desempeño de Puntaje F1 tanto en hombres (0.934 en Español Mozilla y 0.93 en Inglés *PVQC*) como en mujeres (0.783 en Español y 0.7 en Inglés Mozilla).
2. El grupo de los 30 coeficientes cepstrales obtuvieron un desempeño de puntaje F1 tanto en hombres (0.623 en Español, 0.6 en Inglés Mozilla) como en mujeres (0.488 en Español y 0.4 en Inglés Mozilla) mayor al 40 % pero menor al 63 %.
3. En cuanto al cambio de ambientes controlados a no controlados (véase la Tabla 5.19), la diferencia de puntaje F1, aproximada en porcentaje, tanto en hombres como en mujeres de los modelos RF_Tono+*VTL_Train_* Ing y RF_Tono+*VTL_Train_* IngC_ valid_Ing es de 4 % y 22 %, respectivamente. Para el uso de los 30 *MFCC* las diferencias de puntajes F1 aproximadas fueron 6 % en hombres y 19 % en mujeres.

Tabla 5.18: Distribución de las métricas del entrenamiento en Inglés PVQD y validación en Español e Inglés Mozilla.

Algoritmos	Precisión hombres	Exhaustividad hombres	Puntaje F1 Hombres	Precisión mujeres	Exhaustividad mujeres	Puntaje F1 mujeres	Exactitud	Área bajo la curva ROC hombres	Área bajo la curva ROC mujeres
RF_Tono_Train_IngC_valid_Esp	0.988	0.743	0.848	0.52	0.969	0.677	0.8358	0.9550	0.9550
RF_Tono +VTL_Train_IngC_valid_Esp	**0.946**	**0.922**	**0.934**	**0.751**	**0.818**	**0.783**	**0.9221**	**0.9740**	**0.9740**
RF_MFCC20_Train_IngC_valid_Esp	0.901	0.446	0.596	0.302	0.83	0.442	0.4918	0.7310	0.7310
RF_MFCC26_Train_IngC_valid_Esp	0.907	0.489	0.635	0.318	0.827	0.459	0.5317	0.7600	0.7600
RF_MFCC30_Train_IngC_valid_Esp	**0.956**	**0.462**	**0.623**	**0.331**	**0.926**	**0.488**	**0.5821**	**0.7970**	**0.7970**
Algoritmos	**Precisión hombres**	**Exhaustividad hombres**	**Puntaje F1 Hombres**	**Precisión mujeres**	**Exhaustividad mujeres**	**Puntaje F1 mujeres**	**Exactitud**	**Área bajo la curva ROC hombres**	**Área bajo la curva ROC mujeres**
RF_Tono_Train_IngC_valid_Ing	0.988	0.74	0.8	0.52	0.96	0.6	0.793	0.952	0.952
RF_Tono +VTL_Train_IngC_valid_Ing	**0.94**	**0.92**	**0.93**	**0.751**	**0.81**	**0.7**	**0.898**	**0.956**	**0.956**
RF_MFCC20_Train_IngC_valid_Ing	0.90	0.446	0.596	0.302	0.8	0.4	0.532	0.742	0.742
RF_MFCC26_Train_IngC_valid_Ing	0.907	0.48	0.63	0.318	0.8	0.46	0.564	0.762	0.732
RF_MFCC30_Train_IngC_valid_Ing	**0.9**	**0.46**	**0.6**	**0.33**	**0.926**	**0.4**	**0.5**	**0.82**	**0.82**

Tabla 5.19: Diferencias entre el aprendizaje en ambientes no controlados (Español e Inglés) y el aprendizaje en el ambiente controlado en el idioma Inglés.

	RF_Tono+VTL+Train_Ing	RF_Tono+VTL+Train_IngC_valid_Ing	Diferencia	Aproximación en porcentaje
Puntaje F1 Hombres	**0.973**	**0.93**	**0.043**	**4 %**
Puntaje F1 Mujeres	**0.917**	**0.7**	**0.217**	**22 %**
Exactitud	**0.959**	**0.898**	**0.06**	**6 %**
	RF_MFCC30+Train_Esp	**RF_MFCC30_Train_IngC_valid_Esp**	**Diferencia**	**Aproximación en porcentaje**
Puntaje F1 Hombres	**0.99**	**0.934**	**0.056**	**0.06**
Puntaje F1 Mujeres	**0.975**	**0.783**	**0.192**	**19 %**
Exactitud	**0.986**	**0.9221**	**0.0639**	**6.4 %**

5.3.6. Discusión de Resultados del Análisis de los Tres Grupos de Características

En general, el grupo de las características del Tono y Tono-*VTL* obtuvieron los valores más altos en sus métricas tanto en un mismo idioma y ambiente, como en el cambio de los mismos. Por el lado de los *MFCC*, se mostraba una clara tendencia de que a mayor cantidad, el desempeño del clasificador iba mejorando. Sin embargo, son sensibles al cambio de idioma y de ambiente. Si bien esto, puede deberse a que solo se consideró la mediana de los valores de cada coeficiente cepstral, los grupos de características del Tono y el Tono-*VTL* pueden utilizarse para la detección de género sin la necesidad de mayor información. Lo anterior es sumamente útil cuando se busca optimizar los costos computacionales de la extracción de características.

5.4. Primer Caso de Prueba del Reconocimento de Género y Edad por Voz

- **Introducción**

  Aunque los sistemas de reconocimiento de género han avanzado significativamente en los últimos años, todavía enfrentan algunos desafíos, especialmente cuando se trata de distinguir entre voces de adolescentes y adultos, ya que los cambios en la voz durante la pubertad pueden dificultar la precisión del reconocimiento de género. Sin embargo, características como la longitud del tracto vocal pueden brindar indicios si la voz del hablante proviene de hablantes adultos o adolescentes. Esto es de especial interés cuando se analizan idiomas tonales como el Thai, que por sus cualidades, los rangos de frecuencia fundamental entre adolescentes y adultos pueden ser cercanos, complicando su detección.

- **Objetivo**

  El objetivo general de esta metodología consiste en el reconocimiento de género y adultez en las categorías adulto (mayor de veinte años), adulta, adolescente masculino (menor a veinte años) y adolescente femenino en el idioma Thai bajo ambientes no controlados.

  Los objetivos específicos son

  - Entrenar cuatro tipos de algoritmos (Bayes Ingenuo, Regresión Logística, Perceptrón Multicapa y AdaBoost en Bosques Aleatorios) de aprendizaje de máquina para el reconocimiento de género y adultez en el idioma Thai.
  - Explicitar la diferencia estadística entre las longitudes de los tractos vocales entre adultos y adolescentes.

De manera general, esta metodología denominada detección de género y adultez (véase Figura 5.5), se considera la longitud del tracto vocal y la estimación de la mediana del cuarto formante como características discriminantes entre adulto (mayor de veinte años) o a un adolescente (menor a veinte años). Los pasos a seguir son los siguientes:

1 Para cada audio de un idioma de Mozilla Common voice se realiza lo siguiente.

1.1 Detección de actividad de voz.

1.2 Extracción de características:

- Longitud del tracto vocal
- Mediana del cuarto formante.
- Extracción del contorno de tono entre un rango de 75 a 350 Hertz y derivar sus 8 características acústicas.

2 Se realiza la validación cruzada con un valor de $k = 10$.

3 Se entrenan los algoritmos: perceptrón multicapa, árbol de búsqueda J48, regresión logística y bosque aleatorio mejorado con Adaboost con los parámetros establecidos en WEKA.

4 Se evalúan las clasificaciones realizadas en términos de sus métricas: precisión, exhaustividad, puntaje F1 y área bajo la curva ROC y exactitud.

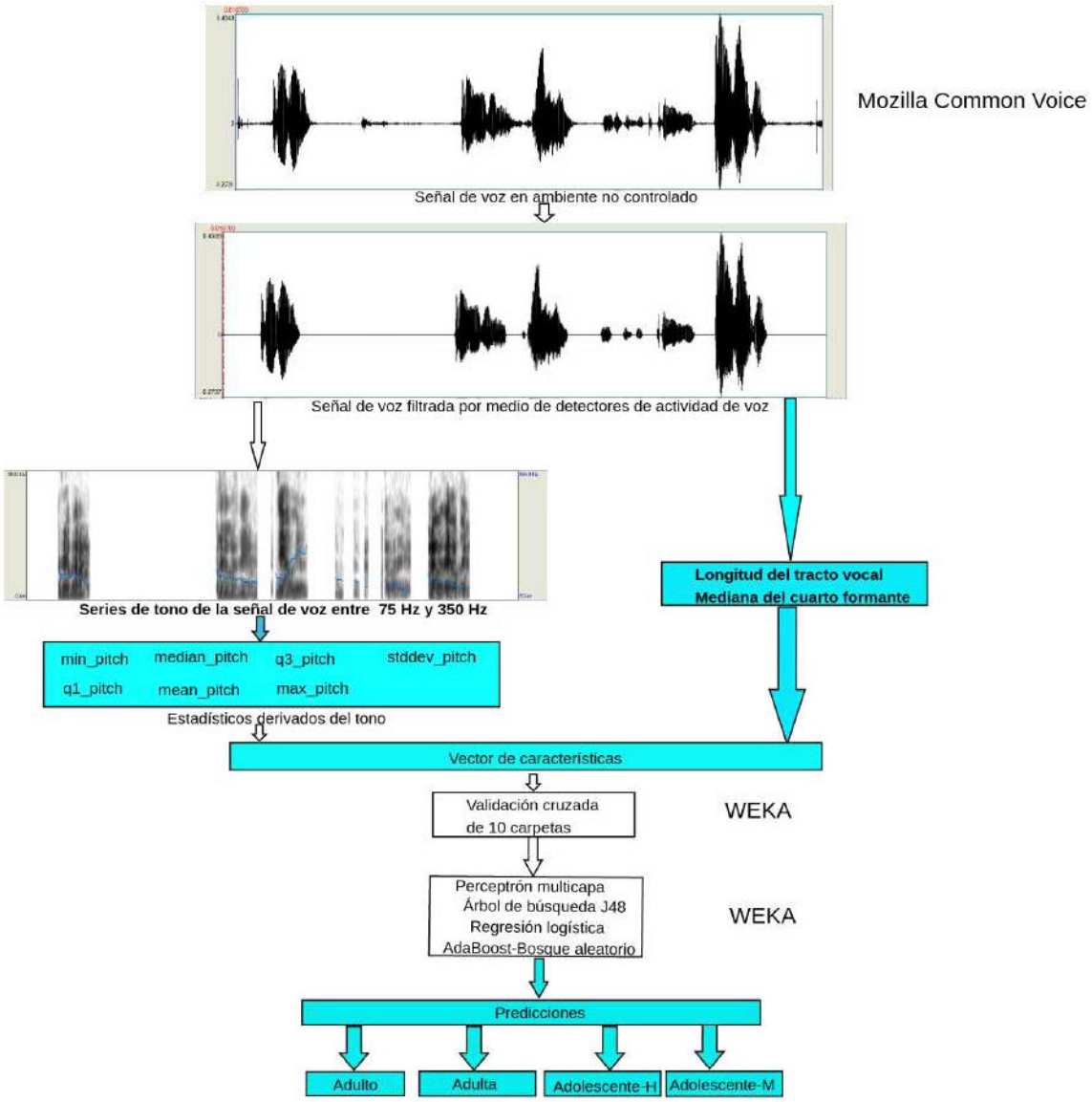


Figura 5.5: Metodología 3 propuesta para el reconocimiento de género y adultez de una voz mediante algoritmos de aprendizaje clásicos.

- **Características de las Pruebas Realizadas**
  El idioma elegido Thai proviene de la base de datos de *Mozilla Common Voice* [13]. La descripción del conjunto de voces así como de las características mediana de F_0 y longitud del tracto vocal pueden verse en las tablas 5.20 y 5.21.

Tabla 5.20: Resumen de la base de voces del idioma Thai.

Audios	**Hombres**	**Mujeres**	**Adolescentes varones**	**Adolescentes mujeres**
79183	50447	28736	2535	1860

Tabla 5.21: Descripción de la mediana de la frecuencia fundamental en grupos de edades en el idioma Thai.

Edad	Género	Cantidad	Media	Desv. est.	Mínimo	Primer cuartil	Mediana	Tercer cuartil	Máximo
Adolescentes	**Mujer**	1860	**227.938806**	28.366302	110.888405	212.118139	**229.043332**	245.188231	328.040808
	Hombre	2535	**148.298229**	34.650882	90.297919	123.181340	**137.528262**	169.824402	298.163486
Veinte	**Mujer**	13160.0	**228.416380**	24.668068	100.959675	214.745609	**230.413001**	243.287255	328.817937
	Hombre	10750.0	**127.058556**	19.838235	81.988102	113.261168	**124.539907**	138.821194	313.748705
Treinta	**Mujer**	4311.0	**219.286653**	23.538171	114.291762	203.271140	**218.404408**	234.655545	317.404218
	Hombre	5650.0	**128.828541**	20.348431	83.809420	115.474715	**125.833195**	138.734796	316.932646
Cuarenta	**Mujer**	465.0	**214.276891**	22.153131	103.107301	200.680915	**213.900682**	226.459181	312.719069
	Hombre	6736.0	**131.329520**	23.186609	79.810511	114.121022	**126.840202**	145.468286	348.764852
Cincuenta	**Mujer**	8888.0	**222.652409**	12.881393	175.246526	213.883126	**221.504950**	230.179576	311.077734
	Hombre	24776.0	**163.107997**	12.808095	83.519518	154.952285	**163.239488**	171.633266	219.709329
Sesenta	**Mujer**	16.0	**209.841920**	15.549202	189.038200	197.586074	**211.209713**	215.588071	244.153215
Ochenta	**Mujer**	36.0	**173.091274**	14.065575	140.735682	162.026860	**172.607926**	186.634686	194.237315

Tabla 5.22: Descripción de la longitud del tracto vocal en grupos de edades y género en el idioma Thai.

Genero	Edad	Cantidad	Media	Desviación estándar	Mínimo	Primer cuartil	Mediana	Tercer cuartil	Máximo
	Veinte	13160.0	**15.632507**	1.299584	12.17	14.7900	**15.45**	16.2000	22.22
	Treinta	4311.0	**15.022011**	0.951319	12.22	14.3900	**14.95**	15.5600	22.09
	Cuarenta	465.0	**15.390000**	0.912983	13.14	14.7400	**15.31**	15.8700	18.91
Mujeres	**Cincuenta**	8888.0	**15.120541**	0.760937	12.76	14.6100	**15.03**	15.5400	19.13
	Sesenta	16.0	**15.124375**	0.760315	13.69	14.7225	**15.02**	15.3950	16.55
	Ochenta	36.0	**15.425556**	0.615743	14.13	14.9825	**15.39**	15.8000	16.78
Mujeres	**Adolescente**	1860.0	**12.024645**	1.356283	8.82	11.1175	**11.88**	12.8200	19.30
	Veinte	10750.0	**17.492662**	1.196312	14.09	16.6700	**17.36**	18.1975	24.17
	Treinta	5650.0	**17.440510**	1.152569	13.31	16.7625	**17.37**	18.1000	23.43
Hombres	**Cuarenta**	6736.0	**17.338337**	1.674399	13.39	16.0700	**17.06**	18.5500	24.16
	Cincuenta	24776.0	**16.562558**	0.933830	13.27	15.9400	**16.50**	17.1000	23.91
Hombres	**Adolescente**	2535.0	**11.796158**	1.538474	8.80	10.5500	**11.66**	12.7600	17.05

- **Resultados** Los resultados de la clasificación de género y adultez de la tercera metodología pueden verse en la Tabla 5.9. Se presentan los resultados más relevantes en forma de listado:
  - El clasificador Bayes Ingenuo obtiene resultados relevantes en la tarea de reconocimiento de género y adultez en idioma tailandés, con una precisión global de alrededor del 80-98 %, mostrando que es una buena opción para esta tarea específica.
  - En general, los resultados son mejores para la clasificación de adultos que para la de adolescentes y, especialmente, de mujeres adolescentes. Esto puede deberse a que la voz de los adultos es más estable y madura, lo que facilita su identificación, mientras que la voz de los adolescentes, en particular las mujeres, es más variable y puede ser más difícil de clasificar.
  - La precisión para la clasificación de mujeres adolescentes es la más baja entre todos los grupos, lo que sugiere que esta es la categoría más difícil de identificar con precisión. Esto podría tener implicaciones en la implementación de sistemas de reconocimiento de voz para aplicaciones

específicas en las que la identificación de género y edad es crítica.

- Los puntajes F1 son razonablemente altos en todos los casos, lo que indica un equilibrio entre la precisión y la exhaustividad. Sin embargo, el puntaje F1 para mujeres adolescentes es significativamente menor que para los otros grupos, lo que indica que hay un mayor desequilibrio entre precisión y exhaustividad para esta categoría en particular.
- Los resultados de la exactitud son bastante altos en general, lo que sugiere que el clasificador es capaz de identificar correctamente la mayoría de los casos. Sin embargo, es importante tener en cuenta que la exactitud sola no es suficiente para evaluar la calidad de un modelo de clasificación, ya que puede ser engañosa en situaciones de desequilibrio de clases.
- El algoritmo de AdaBoost en bosques aleatorios fue quien obtuvo la mayor precisión en las cuatro clases: 98.6 % en adultos, 97.3 % en adultas, 91.1 % en adolescentes hombres y 85.2 % en adolescentes mujeres. Teniendo una exactitud del 97.65 %.

Tabla 5.23: Métricas de los clasificadores usados en el reconocimiento de género y adultez en idioma Thai.

Clasificador	Género	Precisión	Exhaustividad	Puntaje F1	Area ROC	Exactitud
Bayes ingenuo (BI)	Adulto	**0.982**	0.989	**0.985**	0.996	**96.92**
	Adulta	0.971	0.961	0.966	0.993	
	Adolescente hombre	**0.839**	0.82	**0.829**	0.962	
	Adolescente mujer	0.789	0.789	0.789	0.973	
Regresión logística (LR)	Adulto	**0.983**	0.99	**0.986**	0.996	**97.27**
	Adulta	0.97	0.973	0.971	0.995	
	Adolescente hombre	**0.875**	0.825	**0.849**	0.986	
	Adolescente mujer	0.863	0.723	0.786	0.983	
Perceptrón multicapa (MLP)	Adulto	**0.985**	0.99	0.988	0.997	**97.52**
	Adulta	0.972	0.976	0.974	0.996	
	Adolescente hombre	**0.902**	0.838	0.869	0.991	
	Adolescente mujer	0.845	0.777	0.81	0.979	
AdaBoost en bosques aleatorios (RF)	Adulto	**0.986**	0.99	0.988	0.993	**97.65**
	Adulta	0.973	0.978	0.975	0.993	
	Adolescente hombre	**0.911**	0.849	0.879	0.981	
	Adolescente mujer	0.852	0.776	0.812	0.958	

5.5. Segundo caso de prueba para el reconocimento de género y edad por voz

- **Introducción**

  Los resultados de la primer metodología de reconocimiento de género y adultez mostraron que la longitud del tracto vocal puede emplearse como un indicador para saber si una voz proviene de un adulto o de un adolescente. Es necesario destacar que esta característica depende del tipo de frase mencionada y posición de los labios al momento de hablar. No obstante, en diversas tareas relacionadas al procesamiento del habla, se requiere determinar de un rango de edad del hablante. De este modo, el empleo de características más sofisticadas como los coeficientes cepstrales de frecuencias de Mel, facilitaría la identificación de la edad de un hablante.

- **Objetivo**

  El objetivo general de esta metodología consiste en el reconocimiento de género, adultez y rango de edad en décadas mediante el uso de dos tipos de redes neuronales tipo perceptrón multicapa.

  Los objetivos específicos son

  - Entrenar dos tipos de redes de perceptrones multicapa donde la primera se enfocará en el reconocimiento de género y adultez. Mientras que la segunda se enfocará en la detección de edad por décadas.

  La metodología desglosada puede verse en la Figura 5.6.

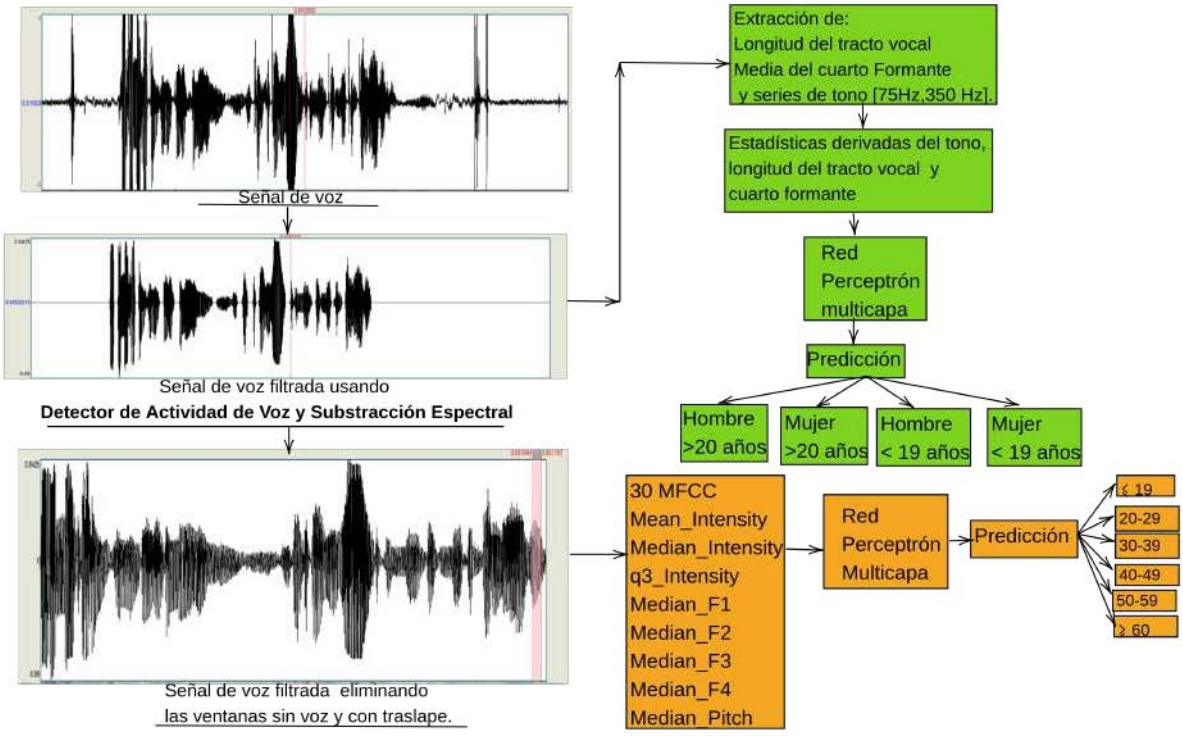


Figura 5.6: Metodología para el reconocimiento de género y edad por voz usando características tonales, formantes y cepstrales.

- **Características de las Pruebas Realizadas**
  Para realizar la experimentación, se utilizó el corpus de Mozilla [13], en los idioma Español con promedio de duración 3.58 seg. y desviación estándar de 1.22 seg. (véase las Tablas 5.25, 5.26, 5.27 y 5.28). Sin embargo, dado que el clasificador es una red de perceptrones multicapa, se optó por combinar los tres conjuntos predefinidos por Mozilla (entrenamiento, prueba y validación) en uno solo. Esto con la finalidad de tener una mayor cantidad de ejemplos de entrenamiento (80 % del total), mientras que se tenga un 10 % para prueba y un 10 % para validar el modelo. Por otro lado, debido a la variabilidad de edades presentadas en el conjunto, se decidió por dividir la edad en 6 categorías (19 años o menos, 20-29, 30-39, 40-49, 50-59, y sesenta años o más). Un desglose de la nomenclatura utilizada puede verse en la Tabla 5.24. Finalmente, cada característica extraída fue winsorizada para poder disminuir la variabilidad presente en los datos.

Tabla 5.24: Categorías consideradas en las diferentes arquitecturas: reconocimiento género adultez, edad, y género-edad (la letra m denota a los hombres y la letra f denota a las mujeres).

Categorías de género-edad (adultos y adolescentes)	**Categorías de edad (en décadas)**	**Categorías de género-edad (en rango de décadas)**
		1) <= 19-m
		2) <= 19-f
		3) 20-29-m
	1) <= 19	4) 20-29-f
1) Mujer mayor o igual a 20 años	2) 20-29	5) 30-39-m
2) Hombre mayor o igual a 20 años	3) 30-39	6) 30-39-f
3) Mujer menor o igual a 19 años	4) 40-49	7) 40-49-m
4) Hombre menor o igual a 19 años	5) 50-59	8) 40-49-f
	6) >=60	9) 50-59-m
		10) 50-59-f
		11) >=60-m
		12) >=60-f

Tabla 5.25: Distribución de la mediana del tono *winsorizada* en el conjunto de voces en Español para el reconocimiento de género y edad (adultos y adolescentes).

Género y edad estimada	**Cantidad de audios**	**Media**	**Desviación estándar**	**Mínimo**	**Primer cuartil**	**Mediana**	**Tercer cuartil**	**Máximo**
Mujer de 20 años o más	53813	**204.77**	19.22	178.23	188.80	**201.80**	220.05	237.58
Hombre de 20 años o más	151133	**122.89**	15.71	98.30	109.90	**123.42**	135.10	147.53
Mujer de 19 años o menos años	3254	**226.25**	20.62	192.22	209.80	**227.74**	243.56	256.19
Hombre de 19 años o menos años	5277	**128.80**	22.43	101.65	110.73	**123.73**	142.05	173.03

Tabla 5.26: Distribución de la longitud del tracto vocal *winsorizado* en el conjunto de voces en Español para el reconocimiento de género y edad (adultos y adolescentes).

Género y edad estimada	Cantidad de audios	Media	Desviación estándar	Mínimo	Primer cuartil	Mediana	Tercer cuartil	Máximo
Mujer de 20 años o más	53813	**17.10**	2.12	**14.61**	15.23	**16.27**	19.46	20.29
Hombre de 20 años o más	151133	**17.27**	1.06	**15.76**	16.28	**17.26**	18.18	18.90
Mujer de 19 años o menos	3254	**12.10**	0.96	**10.61**	11.29	**12.14**	12.89	13.57
Hombre de 19 años o menos	5277	**11.48**	1.18	**9.90**	10.46	**11.31**	12.38	13.53

Tabla 5.27: Distribución de la mediana del tono *winsorizada* en el conjunto de voces en Español para el reconocimiento de edad (en décadas y fusionando el conjunto de entrenamiento, prueba y validación).

Edad	Cantidad de audios	Mínimo	Primer cuartil	Mediana	Media	Tercer cuartil	Máximo
<=19	6534	84.0	119.6	**150.8**	169.7	223.8	340.9
20-29	74722	78.95	128.74	**162.12**	166.44	201.20	341.41
30-39	27585	76.89	117.10	**133.32**	146.74	163.94	309.91
40-49	15991	76.09	114.00	**129.76**	144.72	181.29	318.67
50-59	24327	76.28	99.39	**107.90**	118.67	121.80	347.60
>=60	33056	83.73	123.22	**129.70**	132.56	138.87	318.52

Tabla 5.28: Distribución del primer formante *winsorizado* en el conjunto de voces en Español para el reconocimiento de edad (en décadas y fusionando los conjuntos entrenamiento, prueba y validación).

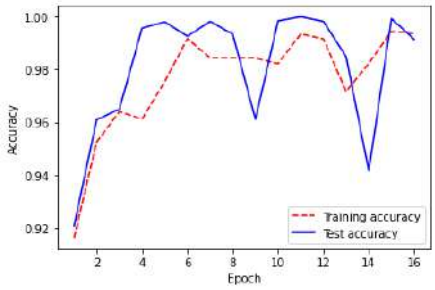
Edad	Cantidad de audios	Mínimo	Primer cuartil	Mediana	Media	Tercer cuartil	Máximo
<=19	6534	240.6	421.0	**462.4**	460.6	499.5	952.2
20-29	74722	207.9	412.0	**446.2**	450.1	485.3	1239.1
30-39	27585	194.8	409.6	**445.1**	447.1	481.7	1080.4
40-49	15991	206.1	400.6	**446.3**	442.7	488.0	1110.5
50-59	24327	239.4	415.5	**444.1**	442.0	469.1	957.1
>=60	33056	185.1	306.5	**327.8**	334.4	352.1	1041.3

- **Resultados**

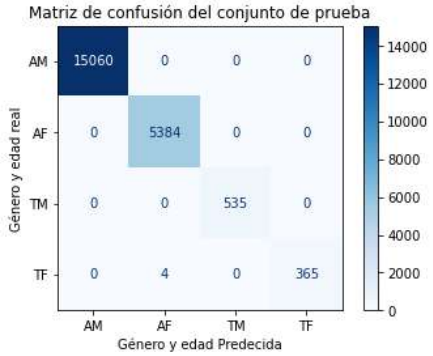
  La red para el reconocimiento de género y adultez fue entrenada con 16 épocas. Las curvas de función de exactitud y pérdida pueden verse en las Figuras 5.7i y 5.7iii. Para validar el modelo, se dividió el conjunto de voces en 80 % entrenamiento, 10 % prueba y 10 % validación. Las matrices de confusión pueden verse en las Figuras 5.7ii y 5.7iv. El desglose de las métricas en el conjunto de prueba y validación se muestran en la Tabla 5.29. Se puede inferir que el alto desempeño de esta red esta asociado a la diferencia estadística significativa entre las longitudes de tracto vocal entre adolescentes y adultos.

  Para el caso de reconocimiento de edad en décadas, se entrenó la red con 400 épocas con un criterio de paro definido detener el entrenamiento si después de 20 épocas, la exactitud no variaba significativamente.

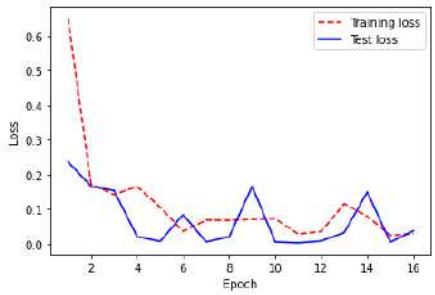
Las curvas de exactitud y función de pérdida pueden verse en las Figuras 5.8i y 5.8iii. Las matrices de confusión pueden verse en las Figuras 5.8iv y 5.8ii. La Tabla 5.30 presenta los resultados de clasificación de voces en rangos de décadas, evaluando la precisión, exhaustividad y puntaje F1 para los conjuntos de prueba y validación. En general, se puede observar que la precisión y la exhaustividad son altas para la mayoría de los conjuntos de edad, con valores que oscilan entre el 0.87 y el 0.95. Además, la precisión y la exhaustividad son particularmente altas para los conjuntos de edad de 50-59 años y mayores de 60 años, con valores que superan el 0.94 en ambos casos. El puntaje F1, que combina la precisión y la exhaustividad, también es alto en la mayoría de los conjuntos de edad, con valores que oscilan entre el 0.87 y el 0.95. En general, los valores de puntaje F1 son ligeramente más bajos para los conjuntos de edad más jóvenes (menores de 30 años) en comparación con los conjuntos de edad mayores. En términos generales, estos resultados sugieren que el modelo de clasificación de voces es efectivo para una amplia gama de edades, con un rendimiento particularmente alto para las personas mayores.
Cabe descatar que el uso de la winsorización en los datos brindó una mejor descripción de los hablantes al atenuar la presencia de datos atípicos. No obstante, es posible que algunos de los objetos cuyas características winsorizadas tendran un valor cercano. Esto es, que existan objetos con los mismo valores de características su diferencia sea computacionalmente despeciable. Por lo anterior, su aprendizaje sera aún más sencillo. Pues, en esencia es el mismo objeto. Esto último solo afectó a lo más al 5 % de todo el conjunto de voces.



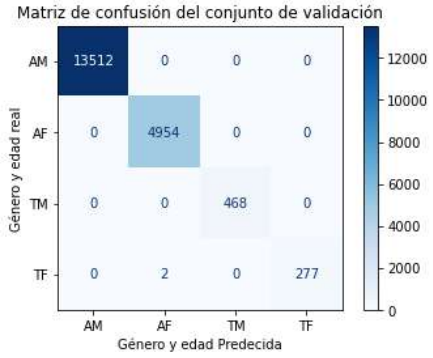
(i) Curva de exactitud.



(ii) Matriz de confusión del conjunto de prueba.

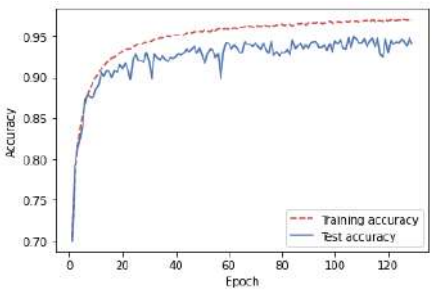


(iii) Función de pérdida.

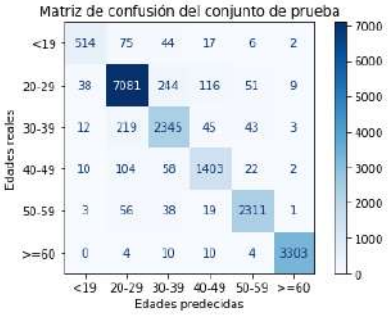


(iv) Matriz de confusión del conjunto de validación.

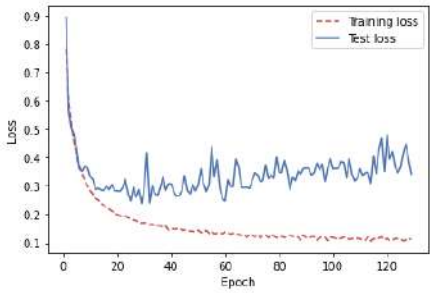
Figura 5.7: Resultados de la red reconocedora de género y adultez.



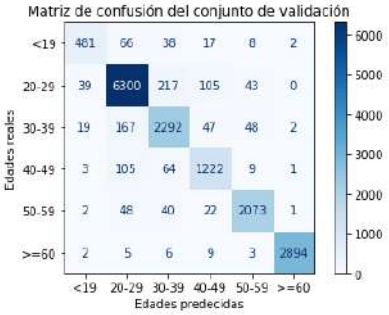
(i) Curva de exactitud.



(ii) Matriz de confusión del conjunto de prueba.



(iii) Función de pérdida.



(iv) Matriz de confusión del conjunto de validación.

Figura 5.8: Resultados de la red reconocedora de edad en décadas.

Tabla 5.29: Métricas del reconocimiento de género y edad (adultez) en el conjunto de prueba y validación.

Conjuntos	Precisión Hombre de 20 años o más	Precisión Mujer de 20 años o más	Precisión Hombre de 19 años o menos	Precisión Mujer de 19 años o menos	Exactitud
Conjunto de prueba	1.0	0.9993	1.0	1.0	0.9998
	Exhaustividad Hombre de 20 años o más	**Exhaustividad Mujer de 20 años o más**	**Exhaustividad Hombre de 19 años o menos**	**Exhaustividad Mujer de 19 años o menos**	
	1.0	1.0	1.0	1.0	
	Puntaje F1 Hombre de 20 años o más	**Puntaje F1 Mujer de 20 años o más**	**Puntaje F1 Hombre de 19 años o menos**	**Puntaje F1 Mujer de 19 años o menos**	
	1.0	0.9996	1.0	1.0	
Conjunto de validación	**Precisión Hombre de 20 años o más**	**Precisión Mujer de 20 años o más**	**Precisión Hombre de 19 años o menos**	**Precisión Mujer de 19 años o menos**	0.9999
	1.0	0.9996	1.0	1.0	
	Exhaustividad Hombre de 20 años o más	**Exhaustividad Mujer de 20 años o más**	**Exhaustividad Hombre de 19 años o menos**	**Exhaustividad Mujer de 19 años o menos**	
	1.0	1.0	1.0	1.0	
	Puntaje F1 Hombre de 20 años o más	**Puntaje F1 Mujer de 20 años o más**	**Puntaje F1 Hombre de 19 años o menos**	**Puntaje F1 Mujer de 19 años o menos**	
	1.0	0.9998	1.0	1.0	

Tabla 5.30: Métricas del reconocimiento de edad (en decadas) en el conjunto de prueba y validación.

Conjuntos	Precisión <=19	Precisión 20-29	Precisión 30-39	Precisión 40-49	Precisión 50-59	Precisión >=60	Exactitud
Conjunto de prueba	0.8908	**0.9392**	0.8562	0.8714	**0.9483**	**0.9949**	**0.9306**
	Exhaustividad <=19	**Exhaustividad 20-29**	**Exhaustividad 30-39**	**Exhaustividad 40-49**	Exhaustividad 50-59	Exhaustividad >=60	
	0.7812	**0.9392**	0.8793	0.8774	**0.9518**	**0.9916**	
	Puntaje F1 <=19	**Puntaje F1 20-29**	**Puntaje F1 30-39**	**Puntaje F1 40-49**	**Puntaje F1 50-59**	**Puntaje F1 >=60**	
	0.8324	**0.9392**	0.8676	0.8744	**0.9501**	**0.9932**	
Conjunto de validación	**Precisión <=19**	**Precisión 20-29**	**Precisión 30-39**	**Precisión 40-49**	**Precisión 50-59**	**Precisión >=60**	**0.9306**
	0.8810	**0.9416**	0.8626	0.8594	**0.9492**	**0.9979**	
	Exhaustividad <=19	**Exhaustividad 20-29**	**Exhaustividad 30-39**	**Exhaustividad 40-49**	**Exhaustividad 50-59**	**Exhaustividad >=60**	
	0.7859	**0.9397**	0.8901	0.8704	**0.9483**	**0.9914**	
	Puntaje F1 <=19	**Puntaje F1 20-29**	**Puntaje F1 30-39**	**Puntaje F1 40-49**	**Puntaje F1 50-59**	**Puntaje F1 >=60**	
	0.8307	**0.9406**	0.8761	0.8648	**0.9487**	**0.9947**	

5.6. Tercer Caso de Prueba del Reconocimento de Género y Edad por Voz

- **Introducción**
  Por los resultados obtenidos en las cinco metodologías anteriores, se eligió diseñar un clasificador que permitiera distinguir el género y la edad en décadas de la persona a través del audio. Para esto se consideraron cuatro grupos de características, a saber, las estadísticas derivadas del tono (primer cuartil y mediana), estadísticas derivadas de la intensidad de la voz (media, mediana, primer y tercer cuartil), medianas de los primeros cuatro formantes del audio, y extracción de la mediana de los primeros 30 *MFCC*. De esta manera, por los resultados anteriores, se espera un desempeño similar a los obtenidos en las metodologías anteriores, pero con la ventaja de obtener más información del hablante.
- **Objetivo** El objetivo general de esta metodología consiste en el reconocimiento de género y edad en décadas por audios en Inglés y Español, mediante el entrenamiento de una red neuronal tipo perceptrón multicapa y la combinación de cuatro tipo de características (tonales, formantes, de intensidad de voz y cepstrales) una red neuronal tipo perceptrón multicapa .
  Los objetivos específicos son
  - Realizar la extracción de características tonales, formantes, de intensidad de voz y cepstrales en un conjunto de voces en Inglés y Español.
  - Realizar el diseño y entrenamiento de la red neuronal.
  - Evaluar el desempeño de la red mediante las métricas de precisión, exhaustividad y medida F1 en un conjunto de prueba y validación.

  La metodología desglosada puede verse en la Figura 5.9.

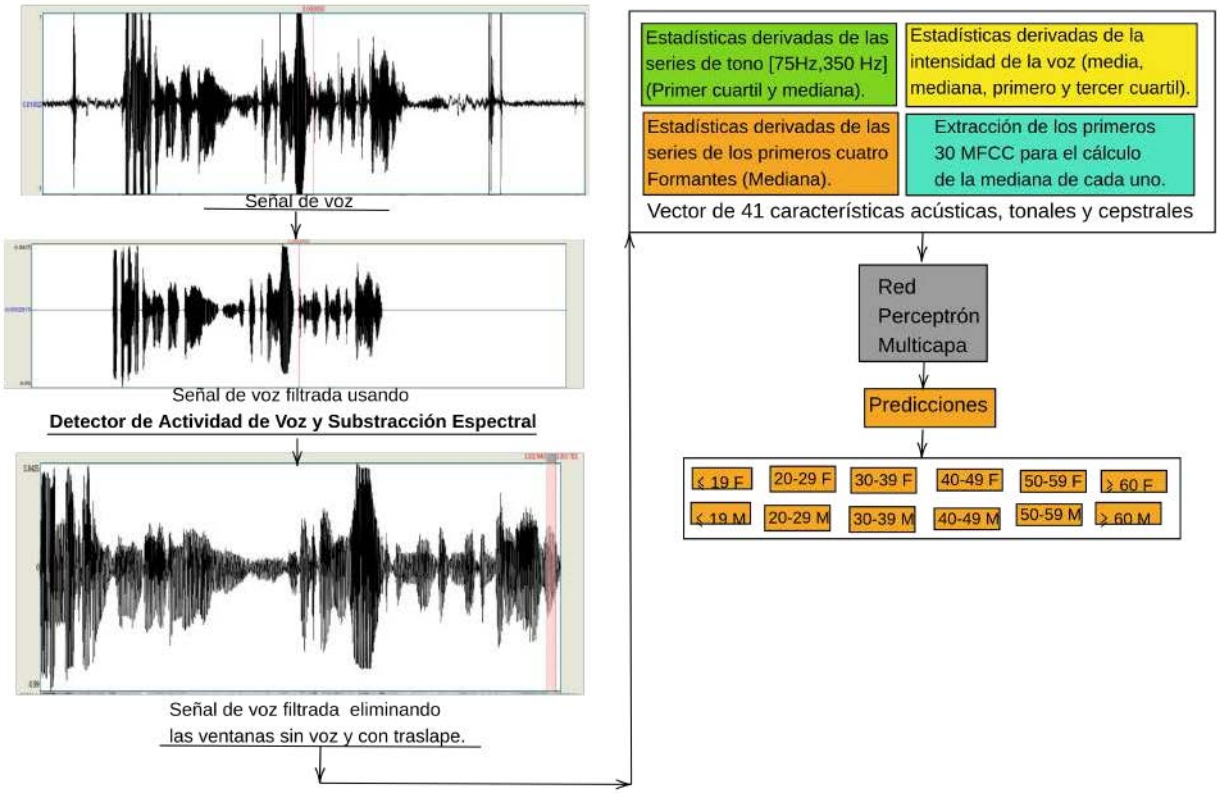


Figura 5.9: Metodología para el reconocimiento de género y edad (en décadas) por voz usando características tonales, formantes y cepstrales.

- **Características de las Pruebas Realizadas**

  Para realizar la experimentación, se utilizó el corpus de Mozilla [13], en los idiomas Inglés con promedio de duración de 2.17 seg. y 1.08 seg. de desviación estándar y Español con promedio de duración 3.58 seg. y desviación estándar de 1.22 seg. (véase las Tablas 5.25, 5.26, 5.27 y 5.28 para el idioma Español y las Tablas 5.31 y 5.32 para la combinación de idiomas Inglés y Español). Sin embargo, dado que el clasificador es una red de perceptrones multicapa, se optó por combinar los tres conjuntos predefinidos en cada conjunto (entrenamiento, prueba y validación). Esto con la finalidad de tener una mayor cantidad de ejemplos de entrenamiento. Por otro lado, debido a la variabilidad de edades presentadas en cada conjunto, se decidió por tres grupos de categorías y *winsorizar* cada característica extraída (véase la Tabla 5.24). De este modo se construyeron los conjuntos de datos Español y Español e Inglés. Mezclar dos idiomas diferentes, en un mismo conjunto de voces para reconocer el género y la edad, es una mejoría a los sistemas de reconocimiento de esta índole por tres razones.

  1. En primer lugar, al utilizar una variedad de idiomas, se pueden recopilar datos de una población más diversa, lo que puede aumentar la precisión y la fiabilidad del sistema. Al tener más datos disponibles, se puede entrenar el modelo para que reconozca una variedad más amplia de patrones y características vocales que pueden ser específicas de un idioma o dialecto determinado.
  2. En segundo lugar, el uso de varios idiomas puede hacer que el sistema sea más adaptable a diferentes contextos y situaciones. En muchos casos, las personas que hablan diferentes idiomas pueden tener diferentes patrones vocales y de habla, lo que puede afectar la precisión del modelo. Al utilizar múltiples idiomas, el sistema puede ser entrenado para ser más flexible y adaptarse a

una variedad de situaciones de habla.

3. En tercer lugar, el uso de varios idiomas puede hacer que el sistema sea más útil en contextos multilingües. En muchos países, la población habla varios idiomas y dialectos, lo que puede dificultar el reconocimiento de la edad por voz si el sistema está diseñado para funcionar solo en un idioma o dialecto. Al utilizar varios idiomas, el sistema puede ser más efectivo para la identificación y verificación de la edad en contextos multilingües.

Tabla 5.31: Distribución de la mediana del tono *winsorizada* en el conjunto de voces en Español e Inglés para el reconocimiento de género y edad en décadas (fusionando los conjuntos de entrenamiento, prueba y validación).

Edades y género	Cantidad	Mínimo	Primer cuartil	Mediana	Media	Tercer cuartil	Máximo
<=19-m	11258	**93.48**	107.03	**119.99**	126.88	138.67	201.24
<=19-f	4369	**175.0**	207.4	**226.2**	224.3	242.4	265.6
20-29-m	76451	**97.94**	114.15	**127.21**	128.39	140.50	167.86
20-29-f	42986	**179.7**	191.9	**203.7**	207.7	220.6	250.6
30-39-m	48856	**92.99**	108.87	**121.40**	123.16	135.92	162.23
30-39-f	14775	**162.9**	186.9	**204.9**	206.4	226.4	251.7
40-49-m	27949	**88.99**	107.22	**119.64**	121.37	133.25	164.90
40-49-f	8585	**149.2**	177.2	**192.3**	194.1	211.7	240.0
50-59-m	31562	**91.0**	100.2	**108.8**	114.0	121.9	160.2
50-59-f	10841	**146.5**	175.2	**203.8**	204.3	232.4	267.2
>=60-m	40660	**105.3**	122.0	**129.3**	131.0	139.5	160.0
>=60-f	3456	**144.6**	171.9	**190.5**	190.5	209.7	236.5

Tabla 5.32: Distribución de la mediana del cuarto formante *winsorizado* en el conjunto de voces en Español e Inglés para el reconocimiento de género y edad en décadas (fusionando los conjuntos de entrenamiento, prueba y validación).

Edades y género	Cantidad	Mínimo	Primer cuartil	Mediana	Media	Tercer cuartil	Máximo
<=19-m	11258	**2740**	3008	**3223**	3211	3420	3645
<=19-f	4369	**2869**	3067	**3182**	3202	3338	3576
20-29-m	76451	**2654**	2981	**3218**	3180	3391	3607
20-29-f	42986	**2509**	2662	**2966**	2983	3216	3763
30-39-m	48856	**2714**	3038	**3234**	3202	3390	3572
30-39-f	14775	**2623**	2944	**3089**	3080	3227	3458
40-49-m	27949	**2742**	3101	**3279**	3241	3415	3588
40-49-f	8585	**2831**	2999	**3136**	3152	3282	3561
50-59-m	31562	**2723**	3055	**3189**	3168	3300	3511
50-59-f	10841	**2806**	3006	**3144**	3159	3297	3584
>=60-m	40660	**3008**	3366	**3469**	3431	3545	3642
>=60-f	3456	**2769**	3016	**3155**	3156	3280	3580

- **Resultados**

  La red para el reconocimiento de género y edad (en décadas) fue entrenada con 100 épocas, con un criterio de paro de máximo 20 épocas sin mejoras significativas en su exactitud. Las curvas de función

de exactitud y pérdida pueden verse en las Figuras 5.10i y 5.10ii. Para validar el modelo, se dividió el conjunto de voces en 80 % entrenamiento, 10 % prueba y 10 % validación. Las matrices de confusión del conjunto de prueba pueden verse en las Tablas 5.33 y 5.34. Para las matrices del conjunto de validación, véase las Tablas 5.35 y 5.36.

En general, Las matrices de confusión de la Tabla 5.33 y de la Tabla 5.34 para el conjunto de prueba muestran que el modelo tiene un rendimiento razonablemente significativo para algunas categorías, pero hay margen de mejora para otras. Al observar la primera fila, que representa a los hombres de 19 años o menos, se observa que hay un alto número de clasificaciones correctas (TP), con 765 muestras clasificadas correctamente (en el caso de la matriz no normalizada). Sin embargo, también hay un número significativo de falsos positivos (FP) en las otras categorías. Esto indica que el modelo puede estar clasificando incorrectamente a algunos hombres mayores como hombres más jóvenes. Por ejemplo, 30 hombres de 20 a 29 años fueron clasificados como menores de 19 años.

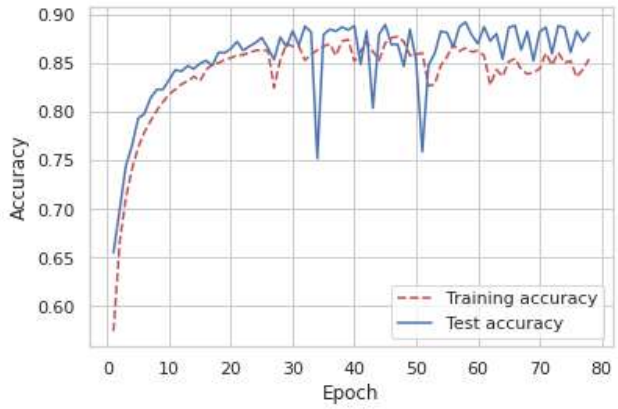
La segunda fila representa a las mujeres de 19 años o menos. En este caso, la mayoría de las clasificaciones son correctas ($TP = 317$), aunque hay algunos falsos positivos en la categoría de mujeres mayores. En general, el modelo parece tener un buen rendimiento en esta categoría.

La tercera fila representa a los hombres de 20 a 29 años. En este caso, hay un alto número de clasificaciones correctas (TP = 7050), pero también hay un número significativo de falsos positivos y falsos negativos en las otras categorías. Por ejemplo, 208 hombres de 19 años o menos fueron clasificados como hombres de 20 a 29 años. Esto indica que el modelo puede tener dificultades para distinguir entre estas dos categorías.

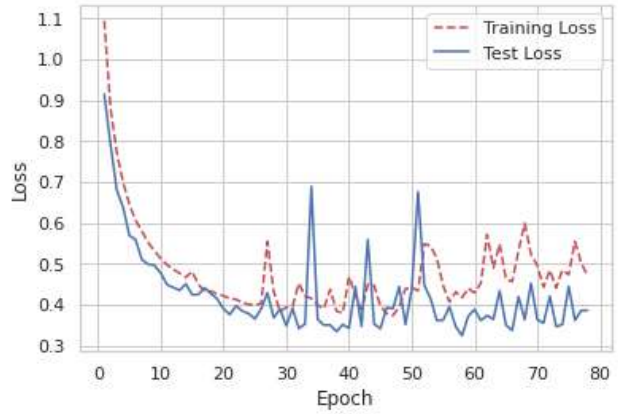
Las filas restantes siguen un patrón similar, con un alto número de clasificaciones correctas en algunas categorías y un rendimiento menos relevante en otras. En general, esta matriz de confusión sugiere que el modelo puede beneficiarse de una mayor capacidad de discriminación entre categorías similares, como hombres de 19 años o menos y hombres de 20 a 29 años.

En la diagonal principal de las matrices de confusión de las Tablas 5.35 y 5.36 del conjunto de validación, se observan los valores verdaderos positivos (en verde) para cada categoría de género y edad, mientras que fuera de la diagonal principal, se presentan las predicciones incorrectas (en rojo). El modelo logró una alta precisión en la detección de género y edad en los hablantes masculinos menores de 20 años y en las hablantes femeninas menores de 19 años, con un total de 685 y 303 verdaderos positivos, respectivamente. Sin embargo, en las categorías de hablantes masculinos entre 20 y 39 años, los valores verdaderos positivos disminuyen, presentando un mayor número de falsos negativos (en amarillo). Además, en estas mismas categorías, se observan varios falsos positivos, lo que indica que el modelo también está prediciendo incorrectamente algunos hablantes femeninos como masculinos.

El desglose de las métricas en el conjunto de prueba y validación se muestran en las Tablas 5.37 y 5.38. El desempeño del modelo para la detección de hombres y su rango de edad es destacable, con valores de precisión, exhaustividad y puntaje F1 altos para la mayoría de los grupos de edad (mayores al 80 % para adultos de 20 años o más). Además, la exactitud general del modelo es del 88.83 % en la prueba y 89.20 % en la validación, lo que sugiere que el modelo tiene una capacidad de clasificación relevante. Sin embargo, es importante destacar que algunos grupos de edad parecen tener un rendimiento menor que otros. Por ejemplo, el grupo de edad de <=19-m años tiene una precisión del 69.23 %, que es más baja que la precisión de cualquier otro grupo de edad. Estos resultados sugieren que el modelo podría beneficiarse de una mayor atención a este grupo de edad para mejorar su rendimiento en el futuro. En el caso del desempeño para la detección de mujeres y su edad, la precisión en las categorías 20-29-f, 30-39-f, 50-59-f y >=60-f es mayor al 80 %, sin embargo las categorías restantes tienen una precisión menor al 78 %. Lo que indica que se requieren una mayor cantidad de ejemplos para mejorar el desempeño en sus reconocimientos.



(i) Curva de exactitud.



(ii) Función de pérdida.

Figura 5.10: Curva de exactitud y función de pérdida de la red de detección de género y edad (en décadas) en el conjunto de prueba.

Tabla 5.33: Matriz de confusión del conjunto de prueba para el reconocimiento de edad y género en el idioma Español e Inglés.

		Predicción del género y edad											
		19-m	**19-f**	**20-29-m**	**20-29-f**	**30-39-m**	**30-39-f**	**40-49-m**	**40-49-f**	**50-59-m**	**50-59-f**	**60-m**	**60-f**
Género y edades reales	**<=19-m**	**765**	3	30	0	38	4	20	5	4	3	3	1
	<= 19-f	3	**317**	0	17	0	7	0	6	0	3	0	3
	20-29-m	208	5	**7050**	7	729	19	378	16	181	17	70	13
	20-29-f	13	67	1	**4174**	0	141	1	124	0	51	0	21
	30-39-m	70	0	314	0	**3916**	1	103	1	81	3	9	2
	30-39-f	7	34	2	71	0	**1196**	1	25	0	10	0	12
	40-49-m	16	0	74	0	125	1	**2372**	0	24	2	3	0
	40-49-f	1	0	6	7	3	7	1	**682**	0	4	0	6
	50-59-m	6	0	69	0	56	0	33	0	**2874**	1	9	0
	50-59-f	5	3	1	10	0	28	2	13	0	**993**	0	6
	>=60-m	8	0	22	0	28	0	20	0	11	0	**3972**	0
	>=60-f	3	2	0	3	0	8	0	6	0	1	0	**272**

Tabla 5.34: Matriz de confusión normalizada del conjunto de prueba para el reconocimiento de edad y género en el idioma Español e Inglés.

		Predicción del género y edad											
		19-m	**19-f**	**20-29-m**	**20-29-f**	**30-39-m**	**30-39-f**	**40-49-m**	**40-49-f**	**50-59-m**	**50-59-f**	**60-m**	**60-f**
Género y edades reales	**<=19-m**	**0.873**	0.003	0.034	0.000	0.043	0.005	0.023	0.006	0.005	0.003	0.003	0.001
	<= 19-f	0.008	**0.890**	0.000	0.048	0.000	0.020	0.000	0.017	0.000	0.008	0.000	0.008
	20-29-m	0.024	0.001	**0.811**	0.001	0.084	0.002	0.043	0.002	0.021	0.002	0.008	0.001
	20-29-f	0.003	0.015	0.000	**0.909**	0.000	0.031	0.000	0.027	0.000	0.011	0.000	0.005
	30-39-m	0.016	0.000	0.070	0.000	**0.870**	0.000	0.023	0.000	0.018	0.001	0.002	0.000
	30-39-f	0.005	0.025	0.001	0.052	0.000	**0.881**	0.001	0.018	0.000	0.007	0.000	0.009
	40-49-m	0.006	0.000	0.028	0.000	0.048	0.000	**0.906**	0.000	0.009	0.001	0.001	0.000
	40-49-f	0.001	0.000	0.008	0.010	0.004	0.010	0.001	**0.951**	0.000	0.006	0.000	0.008
	50-59-m	0.002	0.000	0.023	0.000	0.018	0.000	0.011	0.000	**0.943**	0.000	0.003	0.000
	50-59-f	0.005	0.003	0.001	0.009	0.000	0.026	0.002	0.012	0.000	**0.936**	0.000	0.006
	>=60-m	0.002	0.000	0.005	0.000	0.007	0.000	0.005	0.000	0.003	0.000	**0.978**	0.000
	>=60-f	0.010	0.007	0.000	0.010	0.000	0.027	0.000	0.020	0.000	0.003	0.000	**0.922**

Tabla 5.35: Matriz de confusión del conjunto de validación para el reconocimiento de edad y género en el idioma Español e Inglés.

		Predicción de género y edad											
		19-m	**19-f**	**20-29-m**	**20-29-f**	**30-39-m**	**30-39-f**	**40-49-m**	**40-49-f**	**50-59-m**	**50-59-f**	**60-m**	**60-f**
Género y edades reales	**<=19-m**	**685**	0	28	4	29	9	13	4	6	2	5	2
	<= 19-f	2	**303**	0	16	0	10	0	2	0	1	0	3
	20-29-m	201	2	**6477**	4	635	22	319	23	147	15	62	17
	20-29-f	12	48	2	**3767**	0	117	0	102	0	34	0	15
	30-39-m	55	0	273	0	**3585**	0	105	0	57	3	7	2
	30-39-f	3	37	1	60	0	**1147**	0	18	0	8	0	2
	40-49-m	22	0	83	0	97	1	**1980**	2	31	1	5	0
	40-49-f	2	1	4	12	1	8	2	**607**	0	2	0	5
	50-59-m	9	0	64	0	46	0	31	0	**2507**	0	4	0
	50-59-f	4	9	1	11	3	20	1	11	0	**933**	0	6
	>=60-m	5	0	25	0	21	0	11	0	12	1	**3583**	1
	>=60-f	0	1	1	0	2	3	0	1	0	0	0	**257**

Tabla 5.36: Matriz de confusión normalizada del conjunto de validación para el reconocimiento de edad y género en el idioma Español e Inglés.

		Predicción del género y edad											
		19-m	**19-f**	**20-29-m**	**20-29-f**	**30-39-m**	**30-39-f**	**40-49-m**	**40-49-f**	**50-59-m**	**50-59-f**	**60-m**	**60-f**
Género y edades reales	**<=19-m**	**0.870**	0.000	0.036	0.005	0.037	0.011	0.017	0.005	0.008	0.003	0.006	0.003
	<= 19-f	0.006	**0.899**	0.000	0.047	0.000	0.030	0.000	0.006	0.000	0.003	0.000	0.009
	20-29-m	0.025	0.000	**0.817**	0.001	0.080	0.003	0.040	0.003	0.019	0.002	0.008	0.002
	20-29-f	0.003	0.012	0.000	**0.919**	0.000	0.029	0.000	0.025	0.000	0.008	0.000	0.004
	30-39-m	0.013	0.000	0.067	0.000	**0.877**	0.000	0.026	0.000	0.014	0.001	0.002	0.000
	30-39-f	0.002	0.029	0.001	0.047	0.000	**0.899**	0.000	0.014	0.000	0.006	0.000	0.002
	40-49-m	0.010	0.000	0.037	0.000	0.044	0.000	**0.891**	0.001	0.014	0.000	0.002	0.000
	40-49-f	0.003	0.002	0.006	0.019	0.002	0.012	0.003	**0.943**	0.000	0.003	0.000	0.008
	50-59-m	0.003	0.000	0.024	0.000	0.017	0.000	0.012	0.000	**0.942**	0.000	0.002	0.000
	50-59-f	0.004	0.009	0.001	0.011	0.003	0.020	0.001	0.011	0.000	**0.934**	0.000	0.006
	>=60-m	0.001	0.000	0.007	0.000	0.006	0.000	0.003	0.000	0.003	0.000	**0.979**	0.000
	>=60-f	0.000	0.004	0.004	0.000	0.008	0.011	0.000	0.004	0.000	0.000	0.000	**0.970**

Tabla 5.37: Desglose de las métricas para el género masculino en el conjunto de prueba y validación para el reconocimiento de edad y género en el idioma Español e Inglés.

Conjuntos	Precisión <=19-m	Precisión 20-29-m	Precisión 30-39-m	Precisión 40-49-m	Precisión 50-59-m	Precisión >=60-m	Exactitud
Conjunto de prueba	0.6923	**0.9314**	**0.8000**	**0.8093**	**0.9052**	**0.9769**	**0.8883**
	Exhaustividad <=19-m	**Exhaustividad 20-29-m**	**Exhaustividad 30-39-m**	**Exhaustividad 40-49-m**	Exhaustividad 50-59-m	**Exhaustividad >=60-m**	
	0.87533	**0.8110**	**0.8702**	**0.9064**	**0.9429**	**0.9781**	
	Puntaje F1 <=19-m	**Puntaje F1 20-29-m**	**Puntaje F1 30-39-m**	**Puntaje F1 40-49-m**	**Puntaje F1 50-59-m**	**Puntaje F1 >=60-m**	
	0.7723	**0.8671**	**0.8336**	**0.8551**	**0.9237**	**0.9775**	
Conjunto de validación	**Precisión <=19-m**	**Precisión 20-29-m**	**Precisión 30-39-m**	**Precisión 40-49-m**	**Precisión 50-59-m**	**Precisión >=60-m**	**0.8920**
	0.6850	**0.9307**	**0.8113**	**0.8042**	**0.9083**	**0.9774**	
	Exhaustividad <=19-m	**Exhaustividad 20-29-m**	**Exhaustividad 30-39-m**	**Exhaustividad 40-49-m**	**Exhaustividad 50-59-m**	**Exhaustividad >=60-m**	
	0.8704	**0.8174**	**0.8772**	**0.8911**	**0.9421**	**0.9792**	
	Puntaje F1 <=19-m	**Puntaje F1 20-29-m**	**Puntaje F1 30-39-m**	**Puntaje F1 40-49-m**	**Puntaje F1 50-59-m**	**Puntaje F1 >=60-m**	
	0.7666	**0.8704**	**0.8429**	**0.8454**	**0.9249**	**0.9783**	

Tabla 5.38: Desglose de las métricas para el femenino en el conjunto de prueba y validación para el reconocimiento de edad y género en el idioma Español e Inglés.

Conjuntos	Precisión <=19-f	Precisión 20-29-f	Precisión 30-39-f	Precisión 40-49-f	Precisión 50-59-f	Precisión >=60-f	Exactitud
Conjunto de prueba	0.7355	**0.9732**	**0.8470**	0.7768	**0.9127**	**0.8095**	**0.8883**
	Exhaustividad <=19-f	**Exhaustividad 20-29-f**	**Exhaustividad 30-39-f**	**Exhaustividad 40-49-f**	Exhaustividad 50-59-f	**Exhaustividad >=60-f**	
	0.8904	**0.9088**	**0.8807**	**0.9512**	**0.9359**	**0.9220**	
	Puntaje F1 <=19-f	**Puntaje F1 20-29-f**	**Puntaje F1 30-39-f**	**Puntaje F1 40-49-f**	**Puntaje F1 50-59-f**	**Puntaje F1 >=60-f**	
	0.8056	**0.9399**	**0.8635**	**0.8552**	**0.9242**	**0.8621**	
Conjunto de validación	**Precisión <=19-f**	**Precisión 20-29-f**	**Precisión 30-39-f**	**Precisión 40-49-f**	**Precisión 50-59-f**	**Precisión >=60-f**	**0.8920**
	0.7556	**0.9724**	**0.8579**	**0.7883**	**0.9330**	**0.8290**	
	Exhaustividad <=19-f	**Exhaustividad 20-29-f**	**Exhaustividad 30-39-f**	**Exhaustividad 40-49-f**	**Exhaustividad 50-59-f**	**Exhaustividad >=60-f**	
	0.8991	**0.9195**	**0.8989**	**0.9425**	**0.9339**	**0.9698**	
	Puntaje F1 <=19-f	**Puntaje F1 20-29-f**	**Puntaje F1 30-39-f**	**Puntaje F1 40-49-f**	**Puntaje F1 50-59-f**	**Puntaje F1 >=60-f**	
	0.8211	**0.9452**	**0.8779**	**0.8586**	**0.9335**	**0.8939**	

5.7. Discusión General de los Resultados de las Metodologías

El reconocimento de género mediante el análisis de audios está sujeto a múltiples variables tanto dependientes del hablante como independientes de este. Las metodologías reconocedoras del género mostradas en las secciones anteriores, muestran que existen características relacionadas al tono y a los formantes, que presentan cierta robustez ante el cambio de idioma. Si bien los *MFCC* presentaron un comportamiento similar, su cálculo requiere de un mayor costo computacional. Además, se determinó que la estimación del tracto vocal junto con las características derivadas del tono, son características que no solo pueden emplearse para la detección de género, sino que puede identificar si una voz pertenece a un adolescente o un adulto.

Para realizar una especificación de la edad del hablante, las últimas tres metodologías con sus respectivas redes, a saber, red de tratamiento acústico (para la detección de género y adultez de la voz), red de tratamiento acústico y cepstral (para la detrección de rango de edad en décadas), y la red para modelado de conocimiento acústico (para detección de género y edad en décadas), obtuvieron resultados favorables en cada uno de sus rubros. Si bien, la edad estimada por la voz es aún un problema abierto, los resultados obtenidos indican que las características del tono, formantes, intensidad de voz y *MFCC* pueden emplearse como indicadores para la detección de edad de una persona.

5.8. Comparación con el Estado del Arte

Uno de los aspectos más complicados en la comparación de los resultados reportados en el estado del arte de las diferentes metodologías empleadas, fue que la mayoría de los conjuntos de datos fueron creados para que pudieran brindar la mayor información posible para cada metodología. Sin embargo, si se enfoca esta comparación exclusivamente a:

- La detección de género.
- El conjunto de voces *PVQD* (véase la Tabla 5.12 de la pág. 39).
- El uso de las características derivadas del tono y la estimación del tracto vocal.

Entonces, se tiene que al entrenar tres tipos de algoritmos: J48, Bosques aleatorios o RF, y regresión logística, mejorados por el metaclasificador *AdaBoost* (todos ellos implementados en WEKA). Además, si se consideran tanto las métricas como la condición de validación cruzada de 10 carpetas de la segunda metodología de reconocimiento de género, se tienen los resultados de la Tabla 5.39. La red convolucional presentada en [2] fue considerada, por ser la de mejor desempeño en dicho conjunto de voces. Además, de ser la única en dicho conjunto de voces. Los resultados en la Tabla 5.39 muestran que el desempeño de un clasificador de índole profundo es similar al de clasificadores clásicos, cuando estos utilizan las características del tono y de la estimación del tracto vocal.

Tabla 5.39: Resultados de la exploración de reconocimiento de género por voz usando Adaboost contra redes convolucionales Bensoussan [2].

Algoritmos	Precisión hombres	Exhaustividad hombres	Puntaje F1 Hombres	Precisión mujeres	Exhaustividad mujeres	Puntaje F1 mujeres	Exactitud	Área bajo la curva ROC hombres	Área bajo la curva ROC mujeres
Adaboost-J48_Pitch+VTL_IngC	0.885	0.850	0.867	0.925	0.944	0.934	0.912	0.928	0.928
Adaboost-RF_Pitch+VTL_IngC	**0.903**	**0.840**	**0.870**	**0.921**	**0.954**	**0.937**	**0.916**	**0.941**	**0.941**
Adaboost-LR_Pitch+VTL_IngC	**0.913**	**0.840**	**0.875**	**0.922**	**0.959**	**0.940**	**0.919**	**0.896**	**0.896**
CNN (Bensoussan) Reportado	**0.900**	**0.850**	**0.870**	**0.930**	**0.950**	**0.940**	**0.920**	**0.940**	**0.941**

6 | Conclusiones

6.1. Conclusiones Generales

Los sistemas de reconocimiento de género, género-adultez, y género-edad son una herramienta necesaria en el área de reconocimiento del habla. Ya que permiten extraer diferentes características paralingüísticas útiles para crear perfiles del hablante, y así poder mejorar la interacción humano-computadora.

El presente trabajo realizó seis tipos diferentes de metodologías que exponen (tres para reconocimiento de género y las restantes para reconocimiento de género-adultez y género-edad). Las primeras dos metodologías de reconocimiento de género mostraron el desempeño de las estadísticas derivadas del tono como características robustas para la detección del género en diferentes idiomas. Por lo que es una primera aproximación al estudio del impacto de los idiomas en las áreas de reconocimento de características paralingüísticas de la voz. La tercer metodología del reconocimiento de género mostró la robustez del grupo de características derivadas del tono, la longitud del tracto vocal y el cuarto formante ante el cambio de idioma (Inglés y Español) y ambiente (controlado y no controlado), en comparación con las coeficientes cepstrales de Frecuencias de Mel. Es de suma importancia identificar grupos de características interpretables por seres humanos que puedan brindar información signficativa del hablante a pesar del cambio de idioma y de ambiente de grabación. De este modo, mejoran el desempeño de los algoritmos que entrenan con estas y permiten a los especialistas determinar características de los hablantes.

Las últimas tres metodologías muestran que la detección de género-adultez y género-edad pueden determinarse a partir de la combinación de las características tonales, formantes, de intensidad de voz y cepstrales. De hecho, en el caso de la metodología para reconocimiento de género-adultez, se tiene que la estimación de la longitud del tracto vocal es un rasgo discriminador para identificar voces de adolescentes y adultos. Mientras que los coeficientes cepstrales pueden emplearse como características para la estimación de la edad en décadas. Así, la combinación de los cuatro grupos de características, permiten el reconocimiento de género-edad. Finalmente, se resaltan los siguientes puntos:

- La mediana del tono, tracto vocal y cuarto formante, como características para reconocer género, son robustas ante el cambio de idioma y el cambio de ambiente en idiomas no tonales.
- Características paralinguísticas como la edad, pueden estimarse (bajo ciertas condiciones) a partir de una combinación de características espectrales tonales, formantes e intensidad de la voz.
- La teoría fuente-filtro es un marco teórico que permite simplificar el proceso de producción de voz, permitiendo la extracción de características que describen al hablante. Esto permite el estudio de estimaciones físicas como altura y peso a partir de un audio.

6.2. Objetivos Alcanzados

- El objetivo principal de este proyecto se logró mediante la identificación de las estadísticas derivadas del tono como reconocedoras de género ante diversos idiomas (primera y segunda metodología).
- Además, como adición al cumplimiento del objetivo principal, se estudió la robustez de las características derivadas del tono y del tracto vocal, ante el cambio de idioma y ambiente de grabación (metodología tres). Y se realizó su comparación con el desempeño de los coeficientes cepstrales de frecuencias de Mel. Estos últimos utilizados en el estado del arte.
- Por otro lado, se realizaron tres tipos diferentes de arquitecturas de redes neuronales tipo perceptrones multicapa, encargadas para la detección de género-adultez, edad, y género-edad. Lo que permite una mejor descripción del hablante.Se descata que la última red (género-edad) fue entrenada con voces españolas como inglesas, lo que le permite detectar la edad y género en ambos idiomas.

6.3. Aportaciones

Las aportaciones se listan a continuación:

- Estudio estadístico de las características derivadas del tono para el reconocimiento de género y su impacto en el idioma del hablante.
- Estudio estadístico de la estimación del tracto vocal para el reconocimiento de género y adultez en idiomas tonales y no tonales.
- Estudio de comparativo de desempeño entre dos grupos de características (tonales contra cepstrales) en el reconocimiento de género ante cambio de idioma y de ambiente de grabación.
- Realizar la primera colaboración entre el grupo *Voice Colab AI*[1] y el departamento de ciencias computacionales de CENIDET.
- La propuesta de seis metodologías y evaluación de las mismas
- La propuesta de la longitud del tracto vocal y su evaluación en el reconocimiento de género y edad.

6.4. Trabajo Futuro

Se presenta una lista de los posibles trabajos futuros:

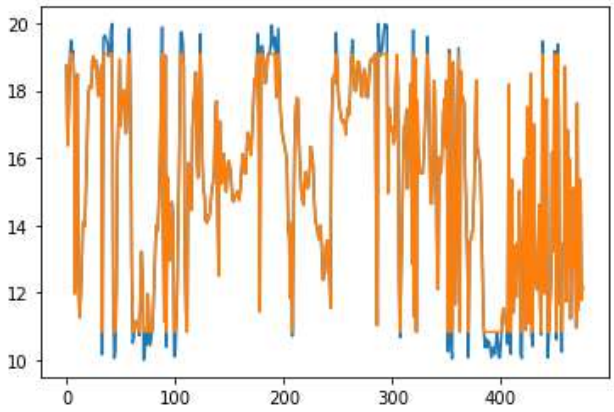
1. **Estudio analítico y estadístico de los formantes en la estimación del tracto vocal, altura y peso para hablantes en ambientes no controlados**
   Planteamiento del problema: Existe una correlación lineal entre la altura de un mamífero, su tamaño (véase [64]) y la estimación de su longitud de tracto vocal. Sin embargo, las estimaciones de este último dependen del tipo de formante utilizado (véase las Figuras 6.1i y 6.1ii). Por lo anterior, la implementación de un sistema que elija el formante adecuado para la estimación de la longitud del tracto vocal, y con este, estimar la altura y peso de un hablante, brindaría una mejor descripción del tipo de hablante. Generando así un perfil afín de este, para facilitar su reconocimento mediante el habla.

   Objetivo: Desarrollar un sistema que estime la estatura y el peso de la persona a partir de la estimación de su tracto vocal por medio de la elección apropiada de alguno de los primeros formantes.

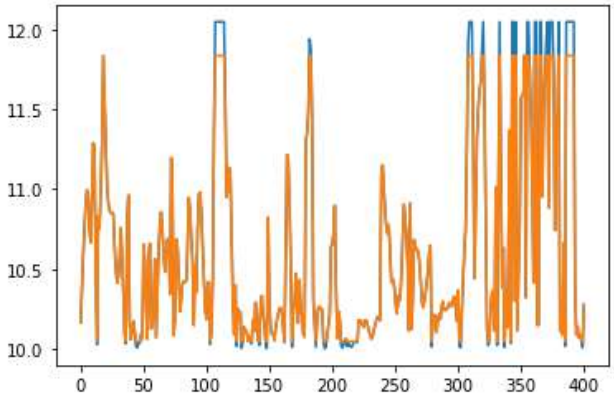
2. **Modelación del pulso glotal con variante fraccionaria del modelo Liljencrants-Fant**
   Planteamiento del problema: El pulso glotal es la onda generada por las cuerdas vocales que viaja a lo largo de tracto vocal. Si el hablante presenta alguna disfonía, esta se ve reflejada en la periodicidad de dicho pulso. Existen modelos matemáticos como el modelo de ecuaciones diferenciales Liljencrants-Fant (véase la Figura 6.2) que modela un pulso glotal. Sin embargo, plantear una variante fraccionaria permitirá estudiar la transición de dicha onda para poder estudiar las variaciones presentes de una voz con alguna patología (como la disfonía) de una voz sana.

   Objetivo: Modelar el pulso glotal mediante una variante fraccionaria del modelo Liljencrants-Fant para el estudio del grado de disfonía de una voz.

[1] https://www.voicecollab.us/team.php



(i) Estimación de la longitud del tracto vocal utilizando el primer formante (curva azul) y el primer formante winsorizado (curva naranja).



(ii) Estimación de la longitud del tracto vocal utilizando el cuarto formante (curva azul) y el cuarto formante winsorizado (curva naranja).

Figura 6.1: Gráficas de estimación de la longitud del tracto vocal según el formante elegido.

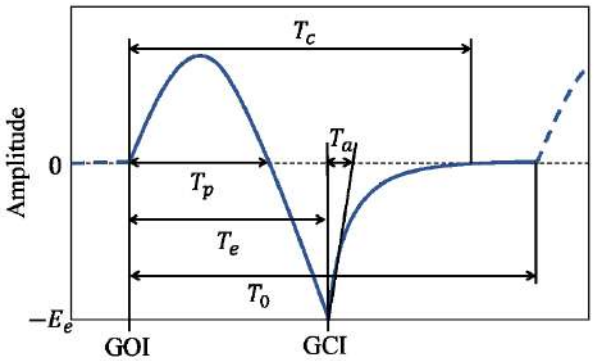


Figura 6.2: Gráfico de la fuente glotal. Fuente: [84] .

6.5. Actividades Académicas Adicionales

La siguientes tablas muestran las actividades académicas adicionales, así como los reconocimientos obtenidos durante esta investigación.

Tabla 6.1: Actividades de difusión, participación en cursos y docencia.

Semestre	Actividad
Primero	Participación como Asistente en la Escuela de Verano en Análisis de Series de Tiempo e Inteligencia Artificial que se llevó a cabo en el Centro de Investigación en Ciencias de la Universidad Autónoma de Morelos (UAEM) del 30 de mayo al 10 de Junio del 2022.
Segundo	Participación en cuatro cursos intersemestrales: Aprendizaje Automático con WEKA, Formalización de soluciones en problemas computacionales, Conversational Workshop, Academic Writing: Tips and Strategies. Todos realizados del 8 al 12 de Agosto del 2022. Charla de divulgación con la temática Importancia del desarrollo de la Investigación en el Área de las Matemáticas y su aplicación en la realidad virtual en la Preparatoria Colegio Teresa de Calcuta en el mes de Octubre del 2022. Una ponencia titulada "Sistemas que detectan el género a través de la voz: introducción, desarrollo y tendencias actuales para el Laboatorio de Sistemas Complejos UAEM llevada a cabo el 23 de Septiembre del 2022. Una conferencia titulada "Sistemas que detectan el género a través de la voz: introducción, desarrollo y tendencias actuales en el Instituto Tecnológico de Zacatepec llevada a cabo el 25 de Octubre del 2022.
Intersemestral	Impartición del curso Intersemestral Análisis de las soluciones en problemas computacionales y su formalización con la colaboración del Prof. José Luís Ramírez Alcantara, impartido del 2 al 6 de Enero del 2023.
Tercero	Participación como Ponente en el Ciclo de Conferencias: "La aplicación de las Matemáticas en la Ingeniería, con el Tema: Modelación Matemática para la Identificación del Sexo Mediante la Voz", el día 23 de mayo del 2023.
Cuarto	Estancia en el Hospital Morsani de la Universidad de Florida del Sur, del 31 de julio del 2023 al 23 de Octubre del 2023, con la Dra. Yael,Bensoussan. Se colaboró con los grupos de investigación en relación con las aplicaciones de la Inteligencia Artificial en el área de la Laringología.

Tabla 6.2: Reconocimientos obtenidos.

Semestre	Actividad
Primero	Miembro activo del *VoiceCollab.ai*, la cual es una colaboración de médicos e investigadores para contribuir a la investigación de laringología con inteligencia artificial y promover la aplicación de modelos de aprendizaje profundo en el campo de la laringología (véase https://www.voicecollab.us/team.php).
Segundo	
Tercero	Primer estudiante del CENIDET galardonado con el Premio Internacional 2023 Eugene N. Myers de la Asociación Americana de Laringología (ALA), por el trabajo *Derivatives of F0 on Voice Gender Recognition by Machine Learning Models Across Different Languages and Noisy Environments* (Derivados de F0 en el reconocimiento del género de la voz por modelos de aprendizaje automático a través de diferentes lenguajes y ambientes con ruido).
Cuarto	

Referencias

[1] Jan Hlavnicka, Roman Cmejla, Jiri Klempir, Evzen Ruzicka, and Jan Rusz. Acoustic tracking of pitch, modal, and subharmonic vibrations of vocal folds in Parkinson's disease and Parkinsonism. *IEEE Access*, 7:150339–150354, 2019.

[2] Yael Bensoussan, Jeremy Pinto, Matthew Crowson, Patrick R Walden, Frank Rudzicz, and Michael Johns III. Deep Learning for Voice Gender Identification: Proof-of-concept for Gender-Affirming Voice Care. *The Laryngoscope*, 131(5):E1611–E1615, 2021.

[3] Damian Kwasny and Daria Hemmerling. Gender and age estimation methods based on speech using deep neural networks. *Sensors*, 21(14), jul 2021.

[4] Anvarjon Tursunov, Mustaqeem, Joon Yeon Choeh, and Soonil Kwon. Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. *Sensors*, 21(17), sep 2021.

[5] Garima Sharma, Kartikeyan Umapathy, and Sridhar Krishnan. Trends in audio signal feature extraction methods. *Applied Acoustics*, 158:107020, 2020.

[6] Lal Zimman. Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Language and Linguistics Compass*, 12(8):1–16, 2018.

[7] Ke Wu and D G Childers. Gender recognition from speech. Part I: Coarse analysis. Technical report, Department of Electrical Engineering, University of Florida, Gainesville Florida, 1991.

[8] D G Childers and Ke Wu. Gender recognition from speech. Part II: Fine analysis. Technical report, 1991.

[9] F. Ertam. An effective gender recognition approach using voice data via deeper LSTM networks. *Applied Acoustics*, 156:351–358, dec 2019.

[10] Mohammed M. Nasef, Amr M. Sauber, and Mohammed M. Nabil. Voice gender recognition under unconstrained environments using self-attention. *Applied Acoustics*, 175:107823, 2021.

[11] Eman H Alkhammash, Myriam Hadjouni, and Ahmed M Elshewey. A Hybrid Ensemble Stacking Model for Gender Voice Recognition Approach. *Electronics*, 11(11), 2022.

[12] Abeer Ali Alnuaim, Mohammed Zakariah, Chitra Shashidhar, Wesam Atef Hatamleh, Hussam Tarazi, Prashant Kumar Shukla, and Rajnish Ratna. Speaker Gender Recognition Based on Deep Neural Networks and ResNet50. *Wireless Communications and Mobile Computing*, 2022:1–13, mar 2022.

[13] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus, url = https://voice., year = 2020. Technical report.

[14] Ricardo Coronado. *Síntesis de voz para el idioma español usando Wavelets*. PhD thesis, Tecnológico Nacional de México / CENIDET, 1999.

[15] Roberto Hernández. *Detección del estado emocional mediante la voz en español de México*. PhD thesis, Tecnológico Nacional de México / CENIDET, 2016.

[16] LaTisha Roberts. *Design and development of a gender and language recognition system.* PhD thesis, Tennessee State University, 2008.

[17] Hassam Sheikh. *Who is speaking? Male or female.* PhD thesis, University of Manchester, 2013.

[18] Hasan Erokyar. *Age and gender recognition for speech applications based on support vector machines.* PhD thesis, University of South Florida, 2014.

[19] Susan Nittrouer, Richard S McGowan, Paul H Milenkovic, and Donna Beehler. Acoustic measurements of men's and women's voices: a study of context effects and covariation. *Journal of Speech, Language, and Hearing Research*, 33(4):761–775, 1990.

[20] Ke Wu and Donald G Childers. Gender recognition from speech. part i: Coarse analysis. *The journal of the Acoustical society of America*, 90(4):1828–1840, 1991.

[21] I Karlsson. Evaluations of acoustic differences between male and females voices: a pilot study, 1992.

[22] Isha Kanani, Heenal Shah, and Sapan H. Mankad. On the performance of cepstral features for voice-based gender recognition. In *Information and Communication Technology for Intelligent Systems*, pages 327–333. Springer Singapore, December 2018.

[23] P Kumar, P Baheti, RK Jha, P Sarmah, and K Sathish. Voice gender detection using gaussian mixture model. *Journal of Network Communications and Emerging Technologies (JNCET)*, 8(4):132–136, 2018.

[24] Pramit Gupta, Somya Goel, and Archana Purwar. A stacked technique for gender recognition through voice. In *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pages 1–3. IEEE, 2018.

[25] Serhat Celil İLERİ, Armağan KARABİNA, and Erdal KILIÇ. Comparison of different normalization techniques on speakers' gender detection. *Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi*, 2(2):1–12, 2018.

[26] Fatih Ertam. An effective gender recognition approach using voice data via deeper lstm networks. *Applied Acoustics*, 156:351–358, 2019.

[27] Rami S. Alkhawaldeh. Dgr: Gender recognition of human speech using one-dimensional conventional neural network. *Scientific Programming*, 2019:7213717, Sep 2019.

[28] Remna R Nair and Bhagya Vijayan. Voice based gender recognition. *International Research Journal of Engineering and Technology*, 6, 2019.

[29] Shivangee Kushwah, Shantanu Singh, Kshitij Vats, and Mrs Varsha Nemade. Gender identification via voice analysis. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pages 746–753, March 2019.

[30] Prasanta Roy, Parabattina Bhagath, and Pradip Das. Gender detection from human voice using tensor analysis. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 211–217, 2020.

[31] Yavuz Selim PINAR, Mucahid Mustafa SARITA, Ilkay CINAR, and Murat KOKLU. Gender determination using voice data. *International Journal of Applied Mathematics Electronics and Computers*, pages 232–235, December 2020.

[32] Lakhan Jasuja, Akhtar Rasool, and Gaurav Hajela. Voice gender recognizer recognition of gender from voice using deep neural networks. In *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, pages 319–324. IEEE, 2020.

[33] Skander Hamdi, Abdelouahab Moussaoui, Mourad Oussalah, and Mohamed Saidi. Gender identification from arabic speech using machine learning. In *International Symposium on Modelling and Implementation of Complex Systems*, pages 149–162. Springer, 2020.

[34] Yael Bensoussan, Jeremy Pinto, Matthew Crowson, Patrick R. Walden, Frank Rudzicz, and Michael Johns III. Deep learning for voice gender identification: Proof-of-concept for gender-affirming voice care. *The Laryngoscope*, 131(5):E1611–E1615, 2021.

[35] Raz Sahar, T. Rao, S. Anuradha, and B. Rao. *Performance Analysis of ML Algorithms to Detect Gender Based on Voice*, pages 163 – 171. 12 2021.

[36] VG Nandan, Sukruth Shivakumar, J Sangeetha, Mukund Pandurang Nayak, and Nishanth SK. A comparative study of deep learning and machine learning approaches in speech emotion and gender recognition system. *NVEO-NATURAL VOLATILES & ESSENTIAL OILS Journal/ NVEO*, pages 12261–12273, 2021.

[37] Mohammed M. Nasef, Amr M. Sauber, and Mohammed M. Nabil. Voice gender recognition under unconstrained environments using self-attention. *Applied Acoustics*, 175:107823, 2021.

[38] Héctor A. Sánchez-Hevia, Roberto Gil-Pita, Manuel Utrilla-Manso, and Manuel Rosa-Zurera. Age group classification and gender recognition from speech with temporal convolutional neural networks. *Multimedia Tools Appl.*, 81(3):3535–3552, jan 2022.

[39] Shankhanil Ghosh, Chhanda Saha, and Naagamani Molakathaala. Neuragen-a low-resource neural network based approach for gender classification. *arXiv preprint arXiv:2203.15253*, 2022.

[40] Antonio Guerrieri, Eleonora Braccili, Federica Sgrò, and Giulio Nicolò Meldolesi. Gender identification in a two-level hierarchical speech emotion recognition system for an italian social robot. *Sensors*, 22(5):1714, 2022.

[41] Abeer Ali Alnuaim, Mohammed Zakariah, Chitra Shashidhar, Wesam Atef Hatamleh, Hussam Tarazi, Prashant Kumar Shukla, and Rajnish Ratna. Speaker gender recognition based on deep neural networks and resnet50. *Wireless Communications and Mobile Computing*, 2022, 2022.

[42] K. Johnson. *Acoustic and Auditory Phonetics*. Acoustic and Auditory Phonetics. Wiley, 2011.

[43] L.R. Rabiner and R.W. Schafer. *Introduction to Digital Speech Processing*. Foundations and Trends in Technology. Lightning Source Incorporated, 2007.

[44] Brad Story. *Mechanisms of Voice Production*, pages 34–58. 04 2015.

[45] Steven L Garrett. *Understanding Acoustics: An Experimentalist's View of Sound and Vibration*. Springer Nature, 2020.

[46] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[47] John Robert Deller, John G. Proakis, and John H. L. Hansen. Discrete-time processing of speech signals. 1993.

[48] Bruno Miguel and Silva Santos. Accurate glottal source estimation and modelling. (June), 2020.

[49] Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2):113–120, 1979.

[50] Mary H. Bellandese. Fundamental frequency and gender identification in standard esophageal and tracheoesophageal speakers. *Journal of Communication Disorders*, 42(2):89–99, mar 2009.

[51] Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic press, 2011.

[52] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. .°'Reilly Media, Inc.", 2022.

[53] Paul Boersma and Vincent van Heuven. Speak and unSpeak with Praat. *Glot International*, 5(9-10):341–347, 2001.

[54] Paul Boersma. Acurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings 17*, 17:97–110, 1993.

[55] G. U. Shagi and S. Aji. A machine learning approach for gender identification using statistical features of pitch in speeches. *Applied Acoustics*, 185:108392, 2022.

[56] Yannick Jadoul, Bill Thompson, and Bart de Boer. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71(2018):1–15, 2018.

[57] K Sreenivasa Rao and KE Manjunath. *Speech recognition using articulatory and excitation source features*. Springer, 2017.

[58] FE White. Fundamentals of acoustics by lawrence e. kinsler, austin r. frey, alan b. coppens, and james v. sanders, 1982.

[59] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A Statistical Model-Based Voice Activity Detection. Technical Report 1, 1999.

[60] Itaru MAEDA. Simple quasi-periodic functions and an inverse power law. *Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics*, 10(1):21–30, 1996.

[61] Kun Il Park and M Park. *Fundamentals of probability and stochastic processes with applications to communications*. Springer, 2018.

[62] Paul Boersma. Acurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings 17*, 17:97–110, 1993.

[63] Sandra de Oliveira Dias. Estimation of the glottal pulse from speech or singing voice. 2012.

[64] W. Fitch and Jay Giedd. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 106:1511–22, 10 1999.

[65] Vani Nair, Pooja Pillai, Anupama Subramanian, Sarah Khalife, and Dr. Madhu Nashipudimath. Voice Feature Extraction for Gender and Emotion Recognition. *International Journal on Recent and Innovation Trends in Computing and Communication*, 9(5), 2021.

[66] International Organization for Standardization. *Acoustics: Normal Equal-loudness-level Contours*. ISO, 2003.

[67] Adam C. Lammert and Shrikanth S. Narayanan. On short-time estimation of vocal tract length from formant frequencies. *PLoS ONE*, 10(7):1–23, 2015.

[68] Brad H. Story, Houri K. Vorperian, Kate Bunton, and Reid B. Durtschi. An age-dependent vocal tract model for males and females based on anatomic measurements. *The Journal of the Acoustical Society of America*, 143(5):3079–3102, 2018.

[69] Ramon. Corretge. Praat vocal toolkit. 2012-2022.

[70] Ramon Corretge. Praat vocal toolkit, 2022.

[71] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[72] Wilfrid J Dixon and Kareb K Yuen. Trimming and winsorization: A review. *Statistische Hefte*, 15(2-3):157–170, 1974.

[73] Ian H Witten, Eibe Frank, Mark A Hall, Christopher J Pal, and MINING DATA. Practical machine learning tools and techniques. In *Data Mining*, volume 2, 2005.

[74] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[75] Patrick R Walden. Perceptual voice qualities database (pvqd): database characteristics. *Journal of Voice*, 36(6):875–e15, 2022.

[76] Md. Sadek Ali. Gender Recognition System Using Speech Signal. *International Journal of Computer Science, Engineering and Information Technology*, 2(1):1–9, 2012.

[77] Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77, 2002.

[78] J Ross Quinlan. Program for machine learning. *C4. 5*, 1993.

[79] J.R. Quinlan. Improved use of continuous attributes in c4.5. *J. Artif. Int. Res.*, 4(1):77–90, 1996.

[80] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1998.

[81] Daniel Jurafsky and James H Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

[82] Isha Kanani, Heenal Shah, and Sapan H. Mankad. On the performance of cepstral features for voice-based gender recognition. In *Smart Innovation, Systems and Technologies*, volume 107, pages 327–333. Springer Science and Business Media Deutschland GmbH, 2019.

[83] Kavita Chachadi and S. R. Nirmala. Gender Recognition from Speech Signal Using 1-D CNN. *Lecture Notes in Networks and Systems*, 237:349–360, 2022.

[84] Kyoko Takahashi and Masato Akagi. Estimation of glottal source waveforms and vocal tract shape for singing voices with wide frequency range. pages 1879–1887, 11 2018.

[85] Ramon. Corretge. Praat vocal toolkit. 2012-2022.

Anexos

A | Síntesis de artículos representativos del 2018 al 2022

A.1. On the Performance of Cepstral Features for Voice-Based Gender Recognition (*Sobre el rendimiento de las características cepstrales para el reconocimiento de género basado en la voz*) [22]

- **Contexto**: Las características cepstrales han sido ampliamente usadas en el reconocimiento del habla. Últimamente han tenido popularidad para el reconocimiento de características paralinguísticas como el género.
- **Relevancia del trabajo**: El análisis de diversos conjuntos de características cepstrales para comparar sus resultados en los problemas de clasificación de género, ampliará nuestro conocimiento sobre que rasgos específicos de la voz presentan una baja variabilidad dentro de un género y una alta variabilidad entre géneros para distintos hablantes.
- **Objetivo de la investigación**: Analizar los desempeños de múltiples conjuntos de características cepstrales de tiempo corto con variación en el número de sus dimensiones en un sistema de reconocimiento de género.
- **Metodología:** Se implementaron dos bases de datos. La primera fue English language speech database for speaker recognition (ELSDR) (*Base de datos de habla en inglés para el reconocimiento de hablantes*) que consistió en 197 audios de entre hombres y mujeres. La segunda fue The Speaker in the Wild (SITW) (*El hablante en lo salvaje*) que consistió de 300 audios en ambientes no controlados. Posteriormente a cada audio se calcularon las siguientes características espectrales (usando encuadres de 20 milisegundos, saltos de 10 ms y 20 coeficientes en cada uno): **Características estáticas**
  - Coeficientes cepstrales en la frecuencia de Mel (MFCC).
  - Coeficientes cepstrales en la frecuencia lineal (LFCC).
  - Coeficientes cepstrales en la freciencia rectangular (RFCC).
  - Coeficientes cepstrales en la freciencia de Mel inversa (IMFCC).

  Características dinámicas
  - Primera derivada del coeficiente cepstral (Δ).
  - Segunda derivada del coeficiente cepstral (Δ^2).

  Como clasificador se implementó una red neuronal a partir de la biblioteca de Keras en python. Finalmente se hicieron multiples combinaciones de los coeficientes y se implementó el clasificador en cada uno de ellos.
- **Resultados**: La tabla de resultados se compone de la siguiente manera (véase tabla A.1). La primer columna muestra las combinaciones de coeficientes que se hicieron para clasificar el género. La segunda y tercera columna muestran la exactitud en el primer y segundo conjunto de datos respectivamente. La cuarta columna muestra la exactitud del sistema si fuere entrenado en el primer conjunto de datos y después puesto a prueba en el segundo conjunto. Finalmente en el caso de la quinta se entrenó en la segunda y se puso a prueba en la primera. Para los casos donde hay un $-$, se debe a que no fueron reportados en el artículo.

Tabla A.1: Resultados de las diferentes combinaciones de coeficientes cepstrales expuesto en [22].

Sistema	**Exactitud en ELSDR**	**Exactitud en SITW**	**Exactitud en ELSDR/SITW**	**Exactitud en SITW/ELSDR**
MFCC (20 características)	55 %	92 %	66 %	70 %
LFCC (20 características)	55 %	93 %	-	-
RFCC (20 características)	57 %	92 %	-	-
IMFCC (20 características)	55 %	91 %	-	-
MFCC + Δ^2 (40 características)	55 %	65 %	65 %	55 %
LFCC + Δ^2 (40 características)	55 %	65 %	-	-
RFCC + Δ^2 (40 características)	55 %	65 %	-	-
IMFCC + Δ^2 (40 características)	55 %	65 %	-	-
MFCC + Δ + Δ^2 (60 características)	55 %	93 %	65 %	80 %
LFCC + Δ + Δ^2 (60 características)	55 %	90 %	-	-
RFCC + Δ + Δ^2 (60 características)	70 %	93 %	-	-
IMFCC + Δ + Δ^2 (60 características)	55 %	93 %	-	-

- **Conclusiones**: Se realizaron y compararon múltiples sistemas de reconocimiento de género usando características cepstrales. Dentro de las de menor desempeño fueron los sistemas de 40 características. No obstante, los resultados mostraron que existe una gran diferencia entre el rendimiento entre los corpus (ELSDRR y STIW) y dentro del corpus (ELSDRR/STIW y STIW/ELSDRR) para esta tarea. Por lo tanto, se necesitan estudiar características que presenten una baja variabilidad dentro de un mismo género y una alta variabilidad entre géneros para distintos hablantes. Además de algoritmos más complejos que permitan extraer este tipos de características y preservar sus resultados entre otros conjuntos de datos.

A.2. An effective gender recognition approach using voice data via deeper LSTM networks (*Un enfoque eficaz de reconocimiento de género utilizando datos de voz mediante redes LSTM más profundas*) [26]

- **Contexto**: La automatización de la detección de género por voz no es una tarea trivial, por lo que se han propuesto multiples formas para resolver este problema.
- **Relevancia del trabajo**: El siguiente trabajo es pionero en utilizar redes profundas de memoria de corto y largo plazo para el problema de detección de género.
- **Objetivo de la investigación**: Diseñar un sistema de reconocimiento de género mediante voz por medio de una red profunda de memoria de corto y largo plazo.
- **Metodología**:
  1. El conjunto de datos públicos proveniente de Kaggle consta de 3168 muestras de audio compuesto por 1584 audios de ambos géneros. Cada audio está asociado 20 características espectrales por audio y su etiqueta de género.
  2. Para la elección de las características relevantes se utilizó el algoritmo Relieff-based
  3. La red neuronal profunda de memoria de corto y largo plazo consiste en dos redes de memoria de corto y largo plazo. Para ambas redes se usaron 100 neuronas ocultas. Finalmente, como función activadora se utilizó softmax.
  4. Se compararon los resultados de la red propuesta con otros 7 clasificadores (Árbol de decisión, discriminante lineal, regresión logística, máquina de vectores de soporte de núcleo lineal, máquina de vectores de soporte de núcleo cuadrático, máquina de vectores de soporte de núcleo gaussiano fino y k vecinos más cercanos. Las métricas utilizadas fueron precisión, sensibilidad, especificidad y media geométrica de estos dos últimos.

- **Resultados**: Los resultados de los clasificadores pueden verse en la siguiente tabla A.2.

Tabla A.2: Evaluaciones de los clasificadores clásicos y el propuesto en [26].

Clasificador	**Exactitud**	**Sensibilidad**	**Especificidad**	**Media geométrica**
Árbol de desición	96.2 %	94.6 %	97.9 %	96.2 %
Discriminante lineal	96.6 %	97.4 %	95.7 %	96.5 %
Regresión logística	96.5 %	96.0 %	96.9 %	96.4 %
máquina de vectores de soporte de núcleo lineal	96.3 %	96.7 %	96.6 %	96.6 %
máquina de vectores de soporte de núcleo cuadrático	97.2 %	97.2 %	97.2 %	97.2 %
máquina de vectores de soporte de núcleo gausiano	97.3 %	96.9 %	97.7 %	97.3 %
K vecinos más cercanos	97.6 %	97.4 %	97.7 %	97.5 %
Modelo propuesto	98.4 %	97.2 %	99.5 %	98.3 %

- **Conclusiones**: La elección de características relevantes favoreció a la obtención de mejores valores de métricas en los múltiples clasificadores. El modelo propuesto obtuvo los valores más altos en cuanto a exactitud, especificidad y media geométrica de ambos. Mientras que el discriminante lineal y los k vecinos más cercanos obtuvieron los valores de sensibilidad más altos.

A.3. DGR: Gender recognition of Human Speech Using One-Dimensional Conventional Neural Network (*DGR: Reconocimiento del género del habla humana mediante una red neuronal convencional unidimensional*) [27]

- **Contexto**: La voz humana contiene información paralingüística (edad, género, tono, emoción, etc.) utilizada en múltiples aplicaciones en el reconocimiento de voz. En particular, la detección del género está sujeta a múltiples complicaciones. Por lo tanto, los clasificadores deben utilizar múltiples características de la voz a manera de que puedan mejorar su rendimiento.
- **Relevancia del trabajo**: El estudio de ciertas características espectrales en la detección de género en diversos clasificadores, brinda una evaluación sobre su desempeño. Además, aporta información sobre que otras combinaciones entre características y clasificadores considerarse para una mejor detección.
- **Objetivo de la investigación**: Realizar un estudio estadístico de 4 características espectrales utilizadas en el reconocimiento de género por voz, para su posterior implementación en múltiples clasificadores de reconocimiento de género por voz.
- **Metodología**: El conjunto de datos utilizado fue construído a partir de voces artificiales que contenían 20 idiomas, y en cada idioma había 16 muestras de voz y cada muestra contenía 8 audios de voces masculinas y femeninas. Las características que se extrajeron a cada audio son las siguientes:
  - Coeficientes cepstrales de frecuencias de mel (MFCC).
  - Características de tipo croma (croma).
  - Contraste espectral (contraste).
  - Características de centroide tonal (tonal).

  Los clasificadores implementados fueron los siguientes:
  - Red bayesiana (BN).
  - Red bayesiana ingenua (NB).
  - Red de perceptrones multicapa (MLP).
  - Red neuronal convolucional profunda (DLnorm).
  - Asignación latente de Dirichlet (LDA).
  - Máquina de soporte vectorial lineal (SL)
  - Máquina de soporte vectorial polinómica (SP).
  - Máquina de soporte vectorial radial (SR).
  - Optmización secuencial mínima para máquinas de soporte vectorial (SMO).

- Regresión logística (L).
- Aprendizaje vago de tipo IBk (IBk).
- Aprendizaje vago de tipo K estrella (K*).
- Optimizador adam (Ada)

Los seleccionadores de características implementados son:

- Optimización por nube de partículas (PSO).
- Búsqueda evolutiva (Evolutionary).
- Búsqueda tipo lobo (Wolf).

Las características a evaluar en cada clasificador fueron la precisión y la exhaustividad.

- **Resultados**: Para el análisis estadístico se calculo la R^2 en las regresiones lineales de cada pareja de características (véase tabla A.3). En cuanto a las métricas de los clasificadores véase la imagen A.1.

Tabla A.3: Resultados del análisis estadístico R^2 de las cuatro estadísticas en [27].

	MFCC	**Croma**	**Contraste**	**Tonal**
MFCC	1	0.3318	0.0319	0.0012
Croma	0.3318	1	0.3470	0.0021
Contraste	0.3198	0.3470	1	$3.2398*10^{-6}$
Tonal	0.0012	0.0021	$3.2398*10^{-6}$	1

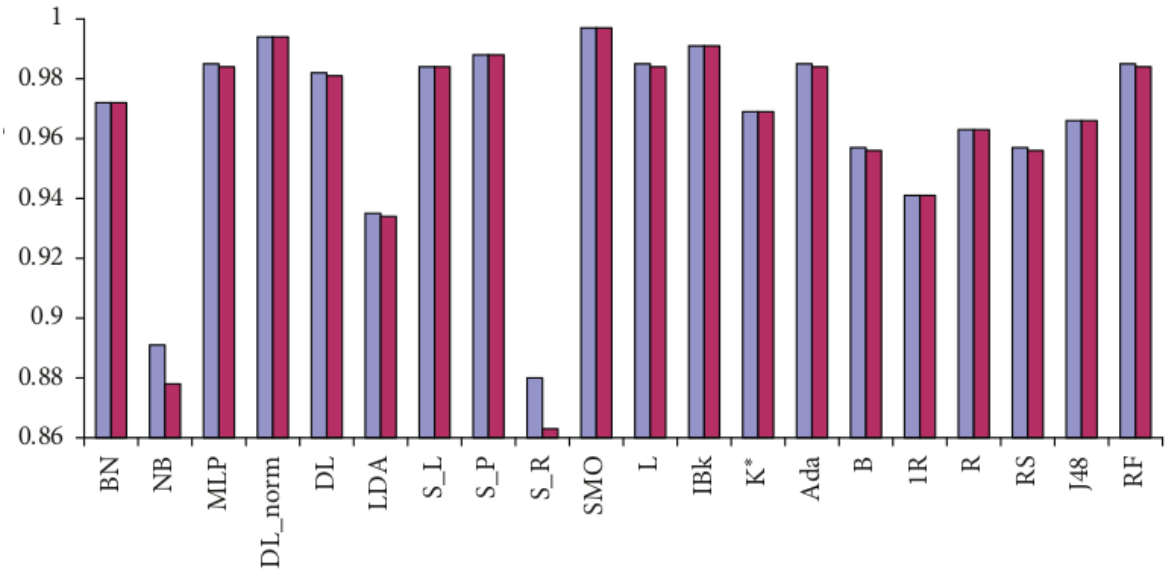


Figura A.1: Gráfico de las métricas de los multiples clasificadores en [27].

- **Conclusiones**: Los algoritmos de aprendizaje profundo (redes neuronales) fueron quienes mostraron un mejor desempeño. Sin embargo, los clasificadores clásicos tuvieron desempeños significativos.

A.4. Voice based gender recognition (*Reconocimiento de género por voz*) [28]

- **Contexto**:
- **Relevancia del trabajo**:
- **Objetivo de la investigación**: Diseñar un sistema de reconocimiento de género mediante la implementación de 4 algoritmos de clasificación, Arboles de decisión (Decision Trees), Potenciación del

gradiente (Gradient Tree Boosting, Bosques aleatorios (Random Forest) y máquinas de vectores de soporte (Support Vector Machine), eligiendo el de mejor desempeño por medio de la métrica de precisión (acuraccy).

- **Metodología**:Se realiza la lectura de los archivos .WAV mediante la paqueteria Scipy-wavefile. Luego, se convierten las amplitudes de los archivos de audio en frecuencias con ventanas de 200 milisegundos, mediante la Transformada de Fourier Discreta (DFT). Las frecuencias son filtradas para contener valores en el rango de la voz humana (20 Hertz a 280 Hertz). Posteriormente, se implementan los algoritmos de clasificación en los vectores obtenidos. Finalmente, se evalúan los resultados.
- **Resultados**: Los resultados pueden verse en la tabla A.4.

Tabla A.4: Resultados de los cuatro clasificadores expuesto en [28].

Algoritmo usado	**Entrenamiento (precisión)**	**Prueba (precisión)**
Árbol de decisión	100 %	86.9 %
Potenciación del gradiente	95.8 %	93.7 %
Bosques aleatorios	98.7 %	89.3 %
Máquinas de soporte vectorial	94 %	90.5 %

- **Conclusiones**: Los algoritmos de clasificación del tipo de máquina de soporte vectorial y potenciación del gradiante demostración la mayor precisión en el conjunto de prueba.

A.5. Deep learning of Voice Gender Identification: Proof-of-concept for Gender-Affirming Voice Care (*Aprendizaje profundo de la identificación de género por voz: prueba de concepto para el cuidado de la voz que afirma el género*) [34]

- **Contexto**: La necesidad de cuidado de la voz que afirme el género ha ido en aumento en la población transgénero en la última década. Actualmente, faltan mediciones objetivas de los resultados del tratamiento para evaluar el éxito de estas intervenciones.
- **Relevancia del trabajo**: Este estudio utiliza modelos de redes neuronales para predecir el género binario a partir de muestras de audio cortas de voces "masculinasz "femeninas". Este el trabajo preliminar es una prueba de concepto para el trabajo futuro para desarrollar una medida de resultado del tratamiento asistido por IA para la afirmación de género mediante el cuidado de la voz.
- **Objetivo de la investigación**:Se diseñó una red convolucional para clasificar el género de una voz mediante el espectrograma obtenido de esta.
- **Metodología**: Doscientas setenta y ocho voces de hablantes masculinos y femeninos de la base de datos de cualidades perceptuales de la voz se utilizaron para entrenar una red neuronal profunda para clasificar las voces como masculinas o femeninas. Cada muestra de audio se asignó a la frecuencia dominio utilizando espectrogramas de Mel. Para optimizar el rendimiento del modelo, se realizó una validación cruzada de 10 veces de todo el conjunto de datos. El conjunto de datos se dividió en 80 % de entrenamiento, 10 % de validación y 10 % de prueba. Las métricas usadas para el estudio fueron la precisión, la exhaustividad y el puntaje F1.
- **Resultados**: Las métricas pueden verse en la tabla A.5.

Tabla A.5: Métricas obtenidas de la red convolucional propuesta en [34].

	Precisión	**Exhaustividad**	**Puntaje F1**
Mujeres	93 %	95 %	94 %
Hombres	90 %	85 %	87 %

- **Conclusiones**: Este estudio de prueba de concepto muestra un rendimiento prometedor para un mayor

desarrollo de una herramienta asistida por IA para proporcionar medidas objetivas de los resultados del tratamiento para el cuidado de la voz de afirmación de género.

A.6. Voice Gender Recognizer-Recognition of Gender from Voice using Deep Neural Networks (*Reconocedor de género de voz- Reconocimiento de género de voz usando redes neuronales profundas*) [32]

- **Contexto**: La motivación de hacer esta investigación es presentar una Voz Modelo de predicción de género para la predicción de Género para un archivo de audio dado.
- **Relevancia del trabajo**: La predicción de género mediante voz es una tarea vital que implica máxima precisión posible. Esta investigación proporciona resultados para establecer una mejor comprensión del aprendizaje profundo y redes neuronales para ayudarlos a identificar el mejor método para predecir el género basado en la voz.
- **Objetivo de investigación**: Construir un modelo para la predicción de género mediante audios de voz basado en redes neuronales profundas.
- **Metodología**: La arquitectura de la red consistió en
  - Capa inicial con 20 unidades
  - Capa interna con 16 unidades y una segunda capa con 8 unidades.
  - Capa exterior con dos unidades de activación con su función sigmoide.

  Para el aprendizaje se calculó la función de pérdida y su derivada con respecto a los parámetros de peso. En su optimización, se implementó el método mini batch y *Stochastic gradient descent*(descenso del gradiente estocástico) (SGD) para el ajuste de hiperparámetros. Además se utilizó el método Adam para la optimización estocástica. Para la activación de la red se utilizó la función ReLU.
- **Resultados**: El modelo de perceptrones multicapa obtuvo mejores resultados, teniendo 96 % de exactitud. No se tuvieron problemas de sobre ajuste (over fitting) pues se usaron las tecnicas de "dropout" y "batch normalisation". Se usó la métrica de exhaustividad obteniendo un 93 %.
- **Conclusiones**: El aprendizaje profundo presenta múltiples ventajas con respecto a las técnicas clásicas de los clasificadores. Sin embargo, requiere de una gran cantidad de datos

A.7. Gender Identification Via Voice Analysis (*Identificación del género mediante el análisis de la voz*) [29]

- **Contexto**: La voz humana es un sonido creado por la vibración de nuestras vocales que se propagan por nuestro tracto vocal. Está compuesta de múltiples elementos y sus características pueden obtenerse a partir del análisis de frecuencias de su señal.
- **Relevancia del trabajo**: El resultado de un clasificador de género de voz está sujeto a las características que se brinden de la voz. Cada característica aporta información relevante para la elección. No obstante, las técnicas para elegir las más relevantes requieren de múltiples procesos. Por lo que usar el criterio del coeficiente de Pearson para descartar las características más similares es una manera rápida para determinar un subconjunto de estas que aporte mayor información al clasificador.
- **Objetivo de la investigación**: Identificar el género usando el análisis acústico de la voz mediante la implementación de 5 clasificadores (regresión logística, árbol de regresión y clasificación, máquina de soporte vectorial, Bosque aleatorio, aumento de gradiente extremo y técnica de ensamblado con los tres últimos).
- **Metodología**: Se consideró una base de datos privada de 3000 audios para ser estudiada. La extracción de 20 características acústicas fue a partir de la biblioteca WarbleR del lenguaje de programación R. Se calculó el coeficiente de Pearson de cada una de las 190 parejas. Posteriormente se descartaron las más afines (sus valoren eran mayores a 0.8) quedando solo 15 de estas. Finalmente, a cada clasificador se les

brindó el conjunto de características seleccionadas.

- **Resultados**: Los resultados de sus clasificadores pueden visualizarse en la tabla A.6.

Tabla A.6: Precisión de los clasificadores en [29].

Clasificador	Precisión
Regresión logística	71 %
Árbol de regresión y clasificación	79 %
Máquina de soporte vectorial	86 %
Bosque aleatorio	88 %
Aumento de gradiente extremo	88 %
Técnica de ensamblado	89 %

- **Conclusiones**: La técnica de ensamblado obtuvo la mayor precisión de todos los clasificadores. No obstante, su diferencia de 1 % no es significativa en comparación con los resultados que suelen obtenerse utilizando aprendizaje profundo.

A.8. 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*) [35]

- **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.
- **Relevancia del trabajo**: La clasificación del género mediante la voz va más allá que solo el estudio de la frecuencia y el tono de una voz humana. Y el elegir las características de la voz que ayuden a determinar el género de una persona es uno de los problemas más complejos del aprendizaje de máquina. Dentro de las características mś populares para la detección de género por voz están los coeficientes cepstrales de frecuencias de Mel. Sin embargo, es posible realizar esta tarea usando solamente estadísticas referentes a las frecuencias de los audios.
- **Objetivo de la investigación**: Este trabajo se examinó el rendimiento de 5 algoritmos clasificadores, los cuales son: bosque aleatorio, árbol de decisión, máquina de soporte vectorial, red neuronal y potenciación del gradiente. En las siguientes métricas: Exactitud (*accuracy*), precisión (*precision*), exhaustividad (*recall*) y puntuación $F1$ ($F1$-*score*).
- **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.
  1. Desviación estándar de la frecuencia
  2. Moda de la frecuencia
  3. Mediana de la frecuencia
  4. Primer cuantil
  5. Tercer cuantil
  6. Curtosis
  7. Centroide espectral
  8. Entropía espectral
  9. Promedio de la frecuencia fundamental medida a través de la señal acústica
  10. Frecuencia fundamental mínima medida a través de la señal acústica
  11. Frecuencia fundamental máxima medida a través de la señal acústica

12. Mínimo de frecuencia dominante medido a través de la señal acústica
13. Promedio de la frecuencia dominante medida a través de la señal acústica
14. Máximo de frecuencia dominante medido a través de la señal acústica
15. Máximo de frecuencia dominante medido a través de la señal acústica
16. 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 A.7.

Tabla A.7: Resultados de los 5 clasificadores expuesto en [35].

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 %

- **Conclusiones**: El método de potenciación del gradiente obtuvo la exactitud más alta seguido de la red neuronal y la máquina de soporte vectorial. Si bien estos valores pueden variar en otras muestras de conjuntos de datos, se pudo demostrar que pueden tenerse resultados significativos utilizando análisis de frecuencias, en vez de los coeficientes cepstrales.

A.9. 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*) [36]

- **Contexto**: El reconocimiento de emociones y género mediante el habla se enfrenta a diversos retos. Uno de estos es el como seleccionar las características adecuadas y que permitan distinguir el genero y la emoción humana ante múltiples escenarios.
- **Relevancia del trabajo**: Comparar las técnicas de aprendizaje profundo con las de aprendizaje de máquina permite visualizar sus ventajas y desventajas a la hora de implementarlas.
- **Objetivo de la investigación**: En este trabajo se examinó el desempeño del aprendizaje por máquina y el aprendizaje profundo en la detección de género (masculino y femenino) y emociones (feliz, enojado, neutral y triste) respectivamente. Para el primer caso se utilizó un modelo de mezclas gausianas mientras que en el segundo utilizó dos conjuntos de clasificadores: por aprendizaje de máquina (árbol de decisión, bosque aleatorio, bayesiano ingenuo, máquina de soporte vectorial, y k vecinos más próximos) y por aprendizaje profundo (dos redes neuronales convolucionales y una red de perceptrones multicapa).
- **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 extraidas por la librería librossa de python.
- **Resultados**: Los resultados de su sistema del clasificador de género pueden verse en la tabla A.8 y para la clasificación de emoción en A.9.

Tabla A.8: Resultados de la clasificación de género en [36].

Resultados del clasificador de género mediante voz usando modelos de mezclas gausianas					
Idioma: Inglés	Hombres: 54 Mujeres: 48 Total: 102		**Idioma: canarés**	Hombres: 50 Mujeres: 50 Total: 100	
Exactitud	Precisión	Exhaustividad	Exactitud	Precisión	Exhaustividad
97.1 %	97.9 %	95.8 %	84 %	85.4 %	82 %

Tabla A.9: Resultados de la clasificación de emociones en [36].

Idioma: Inglés				
Aprendizaje de máquina **Métrica: Exactitud**				
Árboles de desición: 46 %	Bosque aleatorio 57 %	Bayesiano ingenuo 37 %	Máquinas de soporte vectorial 48 %	K-vecinos más cercanos 53 %
Aprendizaje profundo **Métrica: Exactitud**				
Perceptrón multicapa 62 %	Convolucional de una dimensión 70 %		Convolucional de dos dimensiones 97.24 %	
Idioma : Canarés				
Aprendizaje de máquina **Métrica: Exactitud**				
Árboles de desición: 45 %	Bosque aleatorio 55 %	Bayesiano ingenuo 33 %	Máquinas de soporte vectorial 45 %	K-vecinos más cercanos 51 %
Aprendizaje profundo **Métrica: Exactitud**				
Perceptrón multicapa 61.84 %	Convolucional de una dimensión 70.51 %		Convolucional de dos dimensiones 96.77 %	

- **Conclusiones**: Los modelos de aprendizaje profundo tuvieron una mejor exactitud en comparación con los de aprendizaje de máquina. No obstante, los primeros requieren de arquitecturas más sofisticadas para obtener dichos resultados.

A.10. Voice gender recognition under unconstrained environments using self-attention (*Reconocimiento de género de voz en entornos sin restricciones utilizando la atención propia*) [37]

- **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.
- **Relevancia del trabajo**: Los mecanismos de atención son usados en su mayoría en el procesamiento del lenguaje natural (en traducciones principalmente), pero recientemente se han ido implementado en la clasificación de imágenes teniendo rendimientos prometedores. De este modo, mediante la visualización de los coeficientes cepstrales de frecuencias de Mel asociados a una señal de voz, es posible clasificar estas imágenes en los géneros de sus hablantes.
- **Objetivo de la investigación**: Este trabajo presentó dos modelos de reconocimiento de género por voz en ambientes no controlados y basados en atención propia (*self-attention*).
- **Metodología**:
  1. 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.
  2. 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.

3. Se diseñaron dos modelos para este trabajo.
   - El primero consistió de 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 este dos módulos, uno de reducción y el otro de residuales de inicio.

- **Resultados**: La tabla A.10 muestra los resultados de ambos modelos.

Tabla A.10: Resultados de los modelos implentados en [37].

Primer modelo: Atención pura			Número de épocas: 30	
Exactitud	Precisión	Exhaustividad	Puntaje F1	Área bajo ROC
95.11 %	96.07	96.27 %	96.17 %	98.36 %
Segundo modelo: Convolución con atención propia			Número de épocas: 30	
Exactitud	Precisión	Exhaustividad	Puntaje F1	Área bajo ROC
96.23	97.06 %	96.68 %	97.02 %	98.80 %

- **Conclusiones**: Los sistemas de reconocimiento de género mediante la voz basados en atención propia obtuvieron mejores resultados que los implementados en el estado del arte. No obstante el primer modelo tuvo problemas de sobreajuste a los datos, pero el segundo modelo pudo superar este problema gracias a las capas de convolución.

A.11. Gender Detection From Human Voice Using Tensor Analysis (*Detección de género a partir de la voz humana mediante análisis de tensores*) [30]

- **Contexto**: El conocimiento del género de la voz es un gran apoyo para diseñar sistemas de reconocimiento de habla con precisión significativamente mayor a los que no lo implementan.
- **Relevancia del trabajo**: El análisis tensorial es una propuesta diferente a la clásica de modelos de mezclas gausianas usadas en la literatura para la detección de género. Además, no presenta los problemas de mínimos locales como el modelo gausiano.
- **Objetivo**: Diseñar e implementar un sistema de reconocimiento de género por voz mediante el análisis de tensores.
- **Metodología**: Se utilizaron los conjuntos de voces TIMIT-DR1, TIMIT-Mix y SHRUTI (véase la tabla A.11).

Tabla A.11: Distribución de los conjuntos de voces en [30].

Tipo de conjunto de datos	**Conjunto de entrenamiento**		**Conjunto de prueba**	
	Hombres	**Mujeres**	**Hombres**	**Mujeres**
TIMIT-DR1	246	146	34	24
TIMIT Mix	500	500	150	150
SHRUTI	650	650	150	150

Posteriormente, para formar el espacio de vectores de características se utilizaron los coeficientes cepstrales de frecuencias de Mel. El método de los momentos se utilizó para construir la estructura tensorial del espacio de vectores de características para cada género. El método de potencia tensorial se aplicó para calcular los vectores propios de esa estructura tensorial. Las evaluaciones se realizaron mediante la distancia euclidiana.

- **Resultados**: Los resultados se exponen en las tablas A.12 y A.13.

Tabla A.12: Resultados del clasificador basado en análisis tensorial según el tamaño del vector de características propuesto en [30].

Tamaño del vector	**Conjunto de datos**	**Exactitud**		
		Hombre	**Mujer**	**Promedio**
13	Entrenamiento	71.2 %	98.4 %	84.7 %
	Prueba	70.4 %	97.1 %	83.5 %
20	Entrenamiento	92.2 %	76.8 %	84.5 %
	Prueba	95.2 %	72.8 %	84.0 %
26	Entrenamiento	90.8 %	92.4 %	91.2 %
	Prueba	93.36 %	89.82 %	91.59 %

Tabla A.13: Resultados del clasificador basado en análisis tensorial según el número de eigenvectores propuesto en [30].

Número de eigenvectores	**Conjunto de datos**	**Exactitud**		
		Hombre	**Mujer**	**Promedio**
1	Entrenamiento	46.4 %	80.2 %	63.3 %
	Prueba	42.03 %	76.10 %	59.06 %
2	Entrenamiento	72.4 %	86.4 %	79.4 %
	Prueba	75.66 %	84.84 %	84.75 %
3	Entrenamiento	90.0 %	92.4 %	91.2 %
	Prueba	92.92 %	91.15 %	92.03 %
4	Entrenamiento	90.8	92.4	91.2
	Prueba	93.36	89.82	91.59

- **Conclusiones**: El método propuesto alcanza niveles de exactitud cuando el tamaño del vector de características como el número de eigenvectores aumenta.

A.12. Gender identification from arabic speech using machine learning (*Identificación de género a partir del habla árabe mediante aprendizaje automático*) [33]

- **Contexto**: El análisis de reconocimiento de género por voz es un área en desarrollo en los sistemas de reconocimiento del habla. Siendo el idioma una de las condiciones que dificulta su tratamiento.
- **Relevancia del trabajo**: La mayoría de los sistemas de reconocimiento de género están enfocados al idioma inglés. Por lo que cuando son usados en el idioma árabe su desempeño se ve mermado. Esto se debe a las variedades morfológicas de sus letras. Por lo que es necesario desarrollar sistemas en diversos idiomas para entender el problema de detección de género y edad por voz condicionado al idioma.
- **Objetivo**: Diseñar y desarrollar un sistema de reconocimiento de género y estimación de edad basado en la extracción de características cepstrales de frecuencias de Mel del habla árabe, utilizando seis algoritmos de aprendizaje famosos, como árbol de decisiones, bosque aleatorio, K-vecinos más cercanos, máquinas de soporte vectorial, redes neuronales de tipo perceptrones multicapa y bayes ingenuio.
- **Metodología**: El conjunto de voces utilizado fue el corpus Urban Jordan. Este consiste en 12 hablantes nativos de jordano urbano (6 mujeres, 6 hombres). A cada muestra de audio se le extrajeron 13 coeficientes cepstrales. Posteriormente, cada conjunto de características es brindado a los algoritmos.
- **Resultados**: Los resultados de los clasificadores de género pueden verse en la tabla A.14.

Tabla A.14: Resultados de los 5 clasificadores clasificadores [33].

Algortimos de aprendizaje	Exactitud	Precisión	Exhaustividad	Puntaje F1
Red neuronal	96.5 %	92.0 %	91.1 %	92.0 %
Máquina de soporte vectorial	98.5 %	97.9 %	98.9 %	99.1 %
Bosque aleatorio	54.1 %	53.8 %	54.2 %	54.1 %
Árbol de decisión	45.9 %	45.9 %	46.0 %	45.9 %
Bayes ingenuo	43.8 %	40.8 %	42.3 %	43.8 %
K vecinos más cercanos	34.3 %	34.1 %	34.4 %	34.3 %

- **Conclusiones**: La máquina de soporte vectorial y la red neuronal de tipo perceptron multicapa fueron superiores en la determinación de género con precisiones del 98.5 % y 96.5 % respectivamente, mientras que árboles de decisión y bosque aleatorio fueron superiores en la estimación de la edad con precisiones del 95,9 % y 93,0 % respectivamente.

A.13. Voice gender detection using gaussian mixture model (*Detección del género de la voz mediante un modelo de mezcla gaussiana*) [23]

- **Contexto**: La identificación de género se ha implementado en varios sistemas de reconocimiento automático de hablantes y ha demostrado ser de gran importancia. El uso de la identificación de género en la tecnología actual facilita la autenticación e identificación de usuarios en sistemas de alta seguridad.
- **Relevancia del trabajo**: Un sistema de reconocimiento de género puede aplicarse en sistemas más complejos como los relacionados al reconocimiento e identificación del hablante, la anotación multimedia y la indexación del hablante, la anotación en multimedia y el reconocimiento del hablante.
- **Objetivos**: Diseñar e implementar un sistema para la codificación, análisis, síntesis e identificación de género del hablante mediante características cepstrales y usando el clasificador de mezclas gausianas.
- **Metodología**: La base de datos utilizada fue Audio set corpus. Teniendo un total de 50 voces (25 hombres y mujeres), donde a cada una se extrajeron los coeficientes cepstrales de frecuencias de Mel. Posteriormente, se utilizaron estos vectores para construir las distribuciones gausianas (una para hombres y la otra para mujeres).
- **Resultados**: El sistema obtuvo una exactitud de 95 % para mujeres, 76 % para hombres y una exactitud promedio de 86 %.
- **Conclusiones**: Por su simpleza, los modelos gausianos son bastante útiles para detectar el género. Sin embargo, es necesario realizar preprocesamientos a las señales de voz. Con esto, los coeficientes cepstrales pueden brindar una mejor información y así las mezclas gausianas pueden brindar mejores resultados.

A.14. A stacked technique for gender recognition through voice (*Una técnica apilada para el reconocimiento de género a través de la voz*) [24]

- **Contexto**: La clasificación de género tiene aplicaciones como la automatización de los sistemas de vigilancia, analizar las demandas del cliente para la gestión de la tienda entre otros.
- **Relevancia del trabajo**: Utilizar técnicas de ensamblado o apilamiento permite mejorar los resultados individuales de clasificadores de género por voz, obteniendo mejores métricas sin necesidad de sistemas más sofisticados.
- **Objetivos**: Diseñar e implementar un algoritmo de aprendizaje automático apilado para determinar el género utilizando los parámetros acústicos de la muestra de voz y construido a partir de los árboles de clasificación y regresión, máquinas de soporte vectorial (SVM) y red neuronal.
- **Metodología**: El conjunto de voces consistió de múltiples datos acústicos de Harvard-Haskins, VoxForge y Festvox. Obteniendo así 3160 muestras de audio con diferentes duraciones y frecuencias. A cada uno de estos audios se les extrajo las características acústicas por medio de WarbleR (biblioteca del lenguaje

de programación R). Luego, se dividió este conjunto en dos: uno de entrenamiento (2210 muestras) y uno de prueba (950 muestras). Cada uno de los clasificadores y el ensamblado utilizó estos datos.

- **Resultados**: Los resultados pueden verse en la tabla A.15.

Tabla A.15: Resultados del método de ensamblado y de los clasificadores individuales de [24].

Modelo	Exactitud	Tasa de error	Especificidad	Precisión	Exhaustividad	Puntaje F1
CART	95.05 %	4.94 %	93.89 %	94.03 %	96.21 %	95.10 %
Redes neuronales (RN)	95.57 %	4.92 %	96.84 %	96.76 %	94.31 %	95.52 %
Máquinas de soporte vectorial (SVM)	96.01 %	4.0 %	95.78 %	95.80 %	96.21 %	96.0 %
CART-RN-SVM	97.05 %	2.94 %	96.84 %	96.85 %	97.26 %	97.05 %

- **Conclusiones**: Los modelos de ensamble brindan una mejora en las métricas de reconocimiento de género. Esto se debe a que aprovechan las ventajas de cada uno de los clasificadores individuales.

A.15. Comparison of Different Normalization Techniques on Speakers' Gender Detection (*Comparación de diferentes técnicas de normalización en la detección de género de hablantes*)[25]

- **Contexto**: El uso de coeficientes cepstrales de frecuencias de Mel como características para la detección de género ha sido utilizado en múltiples sistemas. Sin embargo, los rangos de las entradas pueden afectar la clasificación. Existen múltiples formas para estandarizar los vectores de características y así tener nuestra información en intervalos compartidos.
- **Relevancia del trabajo**: Las múltiples técnicas para normalizar los vectores de frecuencias de Mel pueden mejorar los resultados de la clasificación. Pero cada normalización difiere en su cálculo por lo que comparar sus desempeños en un mismo clasificador expondrá cuál es el mejor.
- **Objetivos**: Comparar los resultados obtenidos mediante normalizadores de vectores de características de frecuencias cepstrales en un clasificador de máquina de soporte vectorial para la distinción de género por voz.
- **Metodología**: El conjunto de voces es TIMIT. De este se obtuvo un subconjunto audios de 192 hombres y mujeres. A cada audio se les extrajo sus coeficientes cepstrales de frecuencias de Mel. Debido a los diferentes tamaños de vectores obtenidos, se utilizó un análisis de componentes principales para reducir la cantidad de vectores a una estandar. Posteriormente se utilizaron las siguientes normalizaciones:
  - Normalización de media y varianza a corto plazo (STMVN).
  - Tiempo Cepstral Medio y Normalización de Escala (STMSN).
  - Normalización Min-Max.
  - Normalización puntaje Z.
  - Desviación Estándar.
- **Resultados**: Los resultado pueden verse en la figura A.2.

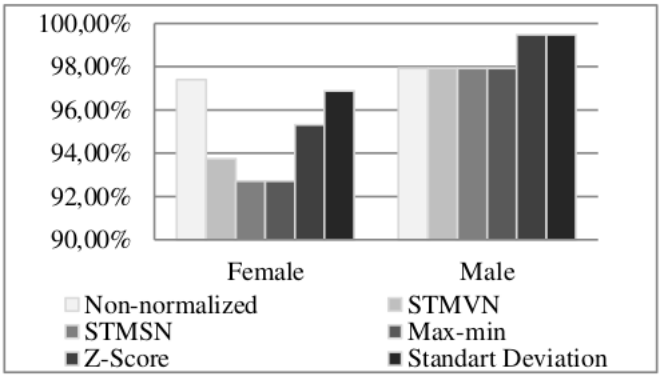


Figura A.2: Gráficas de las diferentes exactitudes según el tipo de normalización propuesto en [25].

- **Conclusiones**: Las normalizaciones están sujetas al tipo de señal con la que se trabaja. En este trabajo, los coeficientes cepstrales provenientes de voces femeninas y no normalizados fueron mejores para el problema de clasificación. Teniendo así una exactitud del 97 %. Sin embargo, para el caso de voces masculinas, la normalización por desviación estándar mejoró la exactitud, teniendo un 99 % de esta, en comparación del 98 % de los no normalizados.

A.16. Gender Determination Using Voice Data (*Determinación del género mediante datos de voz*)[31]

- **Contexto**: El reconocimiento del habla tiene un rol fundamental en la interacción humano computadora. Sin embargo, los factores ambientales y de lenguaje pueden afectar la percepción del sistema de reconocimiento de voz.
- **Relevancia del trabajo**:Los sistemas de reconocimiento de voz con módulos de clasificador de género obtienen mejores resultados. Por lo que indagar en el diseño de estos últimos es un área que ha tenido mucha popularidad estos años.
- **Objetivo principal**: Diseñar e implementar un sistema de reconocimiento de género por voz usando características acústicas y una red neuronal de perceptrones multicapa como clasificador.
- **Metodología**: Se analizó un conjunto de voces de Kaggle con 3168 muestras de audio (1584 hombres y mujeres). Cada muestra de audio se extrajo sus 20 características acústicas usando la librería de WarbleR. En cuanto a la arquitectura de la red, consistió en 3 capas con una capa oculta de 100 neuronas y una de salida con dos neuronas.
- **Resultados**: Los resultados pueden verse en la tabla A.16.

Tabla A.16: Resultados del método de ensamblado y de los clasificadores individuales de [31].

Métrica	Porcentaje
Exactitud	97.9 %
Exhaustividad	98 %
Precisión	97.7 %
Puntaje F1	97.9 %

- **Conclusiones**: Las redes neuronales de tipo perceptrones multicapa brindan un buen desempeño en cuanto a estadísticas. Sin embargo, requiere de 20 parámetros acústicos para obtenerlos. Por lo que cabe la pregunta si existen otras combinaciones de parámetros más pequeñas que puedan obtener los mismos resultados.

A.17. Age group classification and gender recognition from speech with temporal convolutional neural networks (*Clasificación de grupos de edad y reconocimiento de género a partir del habla con redes neuronales convolucionales temporales*) [38]

- **Contexto**: Los sistemas de respuesta de voz interactiva comenzaron a ser utilizados comercialmente en los años 70 por el sistema bancario, con el objetivo de ofrecer los saldos de las cuentas de los clientes. Al principio eran aplicaciones muy cerradas con costes muy elevados. En los años siguientes, la tecnología se desarrolló exponencialmente, logrando mucha más confiabilidad y agregando funcionalidades como reconocimiento de voz, conversión de texto a voz, capacidades de fax e integración de Internet.
- **Relevancia del trabajo**: La automatización de la clasificación de los hablantes en grupos de edad y género es una herramienta indispensable para los sistemas de respuesta de voz interactiva. Pues permite mejorar las métricas de desempeño de trato con el cliente.
- **Objetivo**: Evaluar el desempeño de 18 redes neuronales de los tipos convolucional y convolucional temporal de diversas arquitecturas, para determinar una configuración óptima en el problema de detección de género y rango de edad en sistemas de respuesta de voz interactivos.
- **Metodología**: Se utilizó la Base de datos "Mozilla Common Voice" en el idioma inglés. La cual consiste en 143170 audios de frases con duración de 2 a 5 segundos. Dichos audios tienen una etiqueta que los agrupa por edades de 10 años, 20 años, hasta los 80 años. Por lo anterior, se crearon los grupos Young Male (YM), Young female (YF), Adult Male (AM), Adult Female (AF), Senior Male (SM) y Senior Female (SF). Para cada audio se obtuvieron los coeficientes cepstrales de frecuencias de Mel, los coeficientes delta y delta cuadrada. Posteriormente se utilizó dicha información para cada red neuronal convolucional y convolucional temporal.
- **Resultados**: Los resultados de las 4 arquitecturas con mejores métricas pueden verse en la tabla A.17

Tabla A.17: Resultados de las 4 mejores arquitecturas de redes neuronales propuestas en [38].

Número de red	Tipo	Parámetros	Análisis por bloque			Análisis por muestra de audio		
			Exhaustividad	Precisión	Puntaje F1	Exhaustividad	Precisión	Puntaje F1
4	Red neuronal recurrente convolucional	90092	76 %	81 %	78 %	74 %	81 %	77 %
14	Redes Convolucionales Temporales	84484	66 %	72 %	69 %	73 %	82 %	77 %
18	Red neuronal convolucional	81028	69 %	76 %	73 %	70 %	78 %	74 %

- **Conclusiones**: Los resultados confirman que todos los tipos de redes obtienen resultados significativos para la clasificación de género, pero la combinación de convolucional y convolucional temporal obtuvieron mejores métricas. También se ha estudiado la influencia del número de parámetros libres en el rendimiento del clasificador, demostrando que cuanto mayor es el tamaño de la red (número de parámetros libres), mejor es el rendimiento en ambas aplicaciones. Sin embargo, no vale la pena utilizar más de 50 mil parámetros, independientemente del tipo de configuración de red utilizada.

A.18. NeuraGen-A Low-Resource Neural Network based approach for Gender Classification (*NeuraGen-A Enfoque basado en redes neuronales de bajos recursos para clasificación de género*) [39]

- **Contexto**: La voz humana es la fuente múltiples rasgos.Estos rasgos ayudan a interpretar diversas características asociadas al hablante y el habla. La automatización del reconocimiento de estos rasgos se complica debido a la falta de conjuntos de datos con características específicas (libres de ruido ambiental o con un ruido específico, dentro de un rango de edad, con personas que padecen de enfermedades crónicas de la garganta, etc.). Además, los clasificadores robustos requieren de una enorme cantidad de información para poder mejorar sus métricas de desempeño.
- **Relevancia del trabajo**: Las redes neuronales con arquitecturas de bajos recursos pueden implementarse en múltiples conjuntos de datos no equilibrados (más hombres que mujeres o vice versa), pequeños

(menores a 300 voces) y obtener métricas altas (valores mayores al 50 %).

- **Objetivo**: Diseñar e implementar una red neuronal de bajos recursos que permita realizar la automatización de la detección de género por voz.
- **Metodología**: Los conjuntos de voces utilizados fueron ELSDSR y TIMIT, de los que hemos extraído 8 características del habla. A cada audio se les extrajo 20 coeficientes cepstrales de frecuencias de Mel. La arquitectura propuesta es la siguiente Capa de entrada de forma (1x27), que es la característica de audio vector de audio. Capa oculta 1 de forma (1x25), con función de activación ReLU en cada unidad. Capa oculta 2 de forma (1x10), con función de activación ReLU en cada unidad. Capa oculta 3 de forma (1x5), con función de activación ReLU en cada unidad. Finalmente una capa de salida de 1 unidad, con función de activación sigmoide.
- **Resultados**: Los resultados pueden verse en la tabla A.18

Tabla A.18: Resultados de las métricas de las redes neuronales propuestas en [39].

Conjuto de datos	Exactitud	Pérdida	Precisión	Exhaustividad	Puntaje F1
20 fold CV	90.74 %	17.21 %	96.29 %	86.66 %	91.22 %
10 fold CV	90.74 %	19.60 %	91.66 %	87.99 %	89.79 %
5 fold CV	88.88 %	26.95 %	88.88 %	94.11 %	91.42 %

- **Conclusiones**: La red neuronal propuesta obtuvo resultados significativos en cada una de sus métricas (mayores a 50 %). Por lo que es una primera versión para el diseño de redes neuronales de bajos recursos.

A.19. Gender Identification in a Two-Level Hierarchical Speech Emotion Recognition System for an Italian Social Robot (*Identificación de género en un sistema de reconocimiento de emociones de habla jerárquica de dos niveles para un robot social italiano*) [40]

- **Contexto**: La interacción humano-robot (HRI) aún no se ha resuelto totalmente; sin embargo, se pueden construir robots capaces de percibir las emociones humanas para que puedan interactuar con los humanos de manera adecuada.
- **Relevancia del trabajo**: Proponer un módulo de reconocimiento de género (GR) para identificar el genero del hablante.
- **Objetivo**: Desarrollar un sistema *Speech Emotion Recognition* (SER) en robots sociales para el monitoreo de pacientes hospitalizados y residentes en el hogar.
- **Metodología**: El proceso de detección de emociones por género consta de tres módulos. Primero se detecta la voz, enseguida se detecta el género de la voz para terminar con la detección de las emociones. Se estudiaron las características cepstrales de Frecuencias de Mel (MFCC) y los centroides de sub-bandas espectrales y su combinación. Para el entrenamiento se utilizó la base de datos emocional italiana EMOVO.
- **Resultados**: Los resultados de los clasificadores pueden verse en la Tabla A.19.

Tabla A.19: Métricas del módulo de reconocimiento de género para múltiples combinaciones de parámetros espectrales con y sin detector de voz propuesto en [40].

Características	NO VAD	VAD
SSC	89.8 %	97.8 %
MFCC	53.4 %	90.6 %
MFC + SSC	72.4 %	94 %

- **Conclusiones**: El sistema propuesto escucha continuamente el entorno proporcionando la información de género del hablante con alta precisión. Sin embargo, los tiempos de procesamiento de los módulos desarrollados son muy variables según la duración y la complejidad de las declaraciones pronunciadas.

A.20. Speaker Gender Recognition Based on Deep Neural Networks and ResNet50 (*Reconocimiento de género del hablante basado en redes neuronales profundas y ResNet50*) [41]

- **Contexto**: Debido a la gran variedad de conjuntos de datos y conjuntos de características para la clasificación es complicado que los algoritmos de reconocimiento obtengan buenos resultados. La categorización de las clases según el género puede ayudar a disminuir el impacto de la variabilidad de género en las características recuperadas.
- **Relevancia del trabajo**: Estudio de rendimiento del ajuste fino de ResNet34 y ResNet50 pre-entrenados en espectrogramas de audio para determinar si los espectrogramas brindan suficiente detalle para una clasificación de audio de género precisa. Además, se comparó el rendimiento de varios modelos clásicos de aprendizaje profundo y redes neuronales profundas entrenado en funciones de voz.
- **Objetivo**: Aprovechar la información incluida en los datos de voz sin requerir una intervención manual significativa eligiendo la información esencial de los espectrogramas de voz para realizar la detección de género.
- **Metodología**: Se utilizó una sub-base de 6995 audios de hombres y 5662 audios de mujeres del sitio web Common Voice (`https://commonvoice.mozilla.org/es`)
- **Resultados**: Los resultados de los clasificadores pueden verse en las tablas A.20 y A.21.

Tabla A.20: Clasificadores propuestos en [41].

Características	Modelo	Precisión
FCC, espectrograma de Mel, Chroma STFT, Tonnetz, contraste espectral	DNN	95.97 %
	MLP	95.81 %
	Clasificador K-vecinos	95.10 %
	Clasificador de Bosque Aleatorio	94.23 %
	Kernel SVC RBF	93.92 %
	SVC	91.63 %
	Clasificador Ada boost	90.13 %
	Clasificador de árbol de decisión	88.70 %
	análisis de discriminante cuadrático	77.33 %
	NB Gaussiano	72.27 %

Tabla A.21: Métricas de los clasificadores propuestos en [41].

Modelo	Exactitud	Exhaustividad	Precisión
ResNet34	97.94 %	98.32 %	98.04 %
ResNet50	98.57 %	99.02 %	98.47 %

- **Conclusiones**: Las redes neuronales propuestas obtuvieron las métricas de exactitud, exhaustividad y precisión superiores al 90 %.

B | Diagrama cajas y bigotes de las medianas del tono

Gráficas del diagrama de bigotes por idiomas y edades.

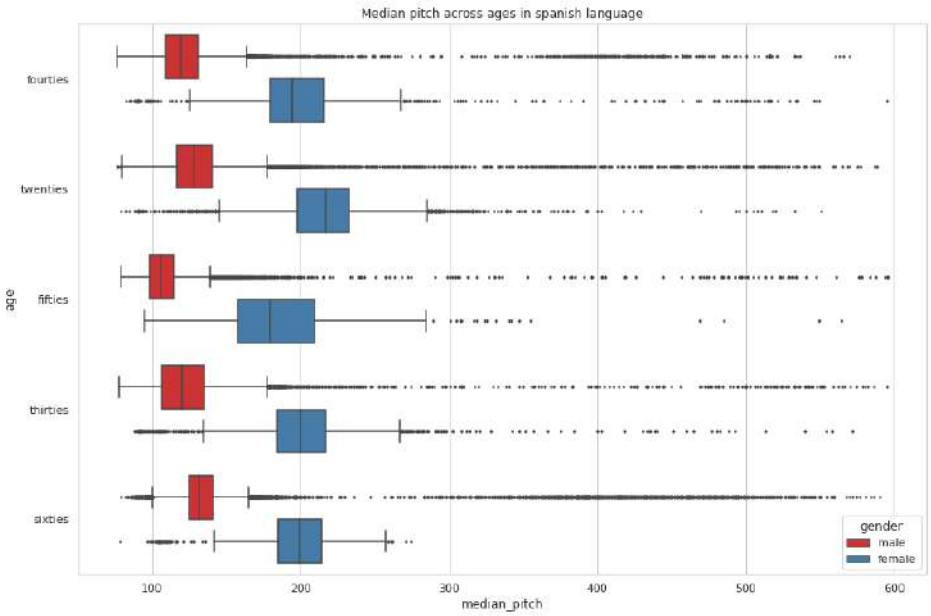


Figura B.1: La mediana del tono en los grupos de edad para el idioma Español.

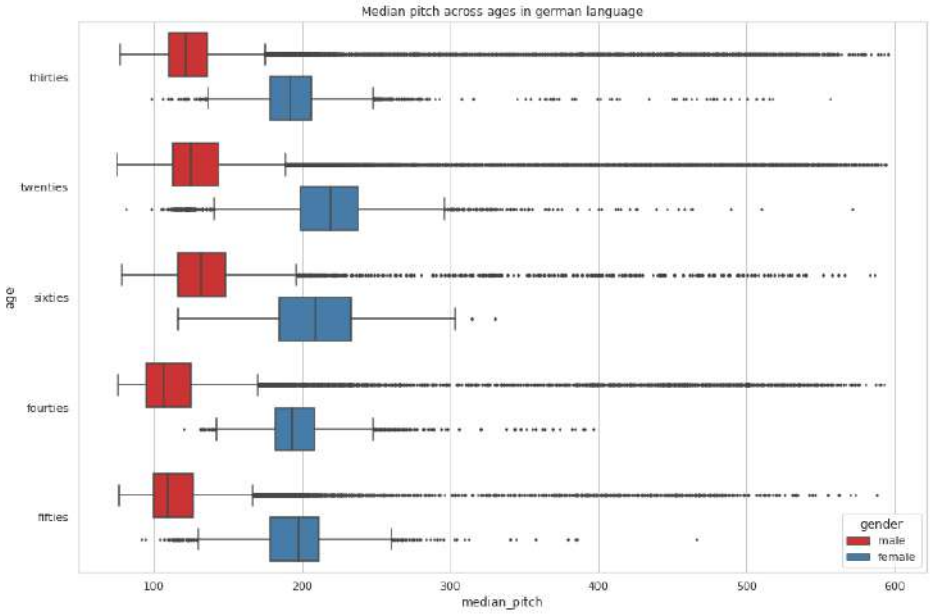


Figura B.2: La mediana del tono en los grupos de edad para el idioma Alemán.

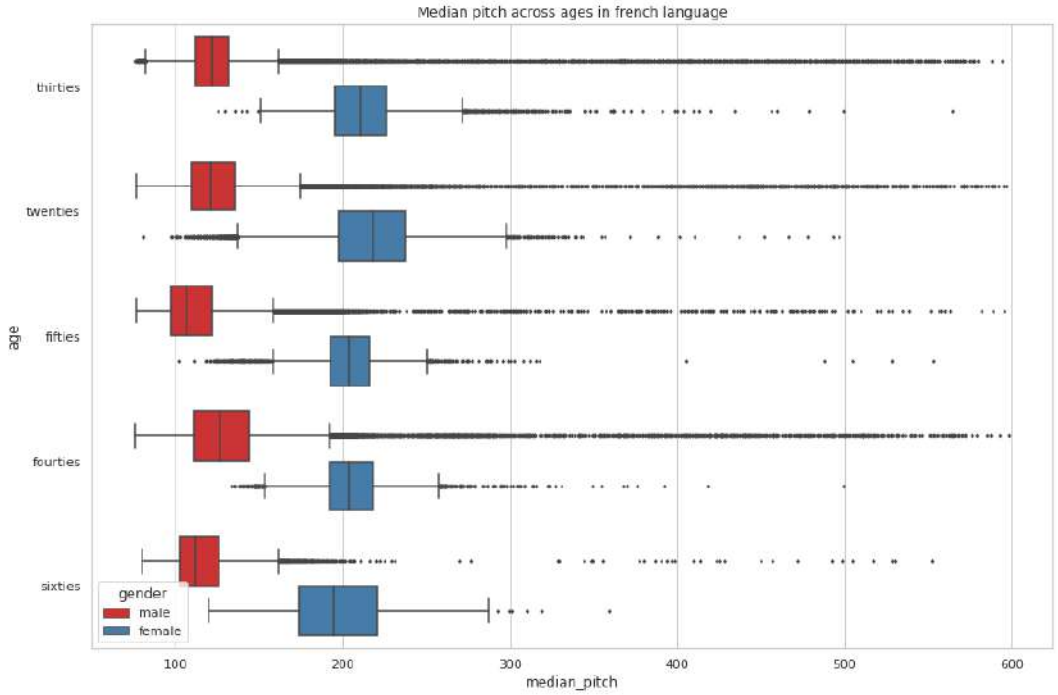


Figura B.3: La mediana del tono en los grupos de edad para el idioma Francés.

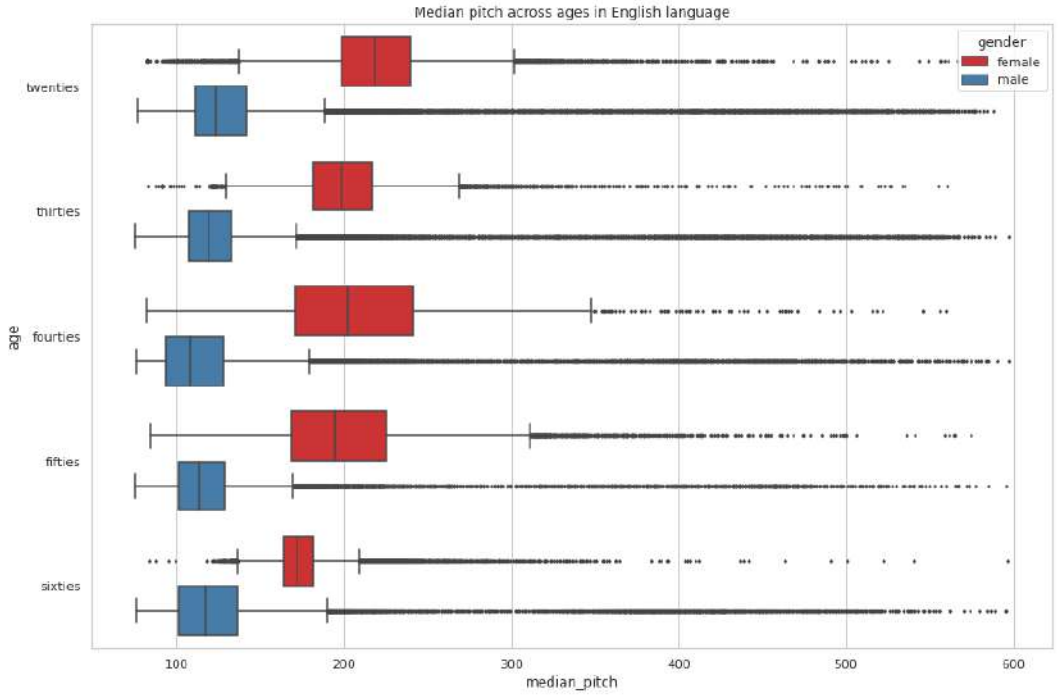


Figura B.4: La mediana del tono en los grupos de edad para el idioma Inglés.

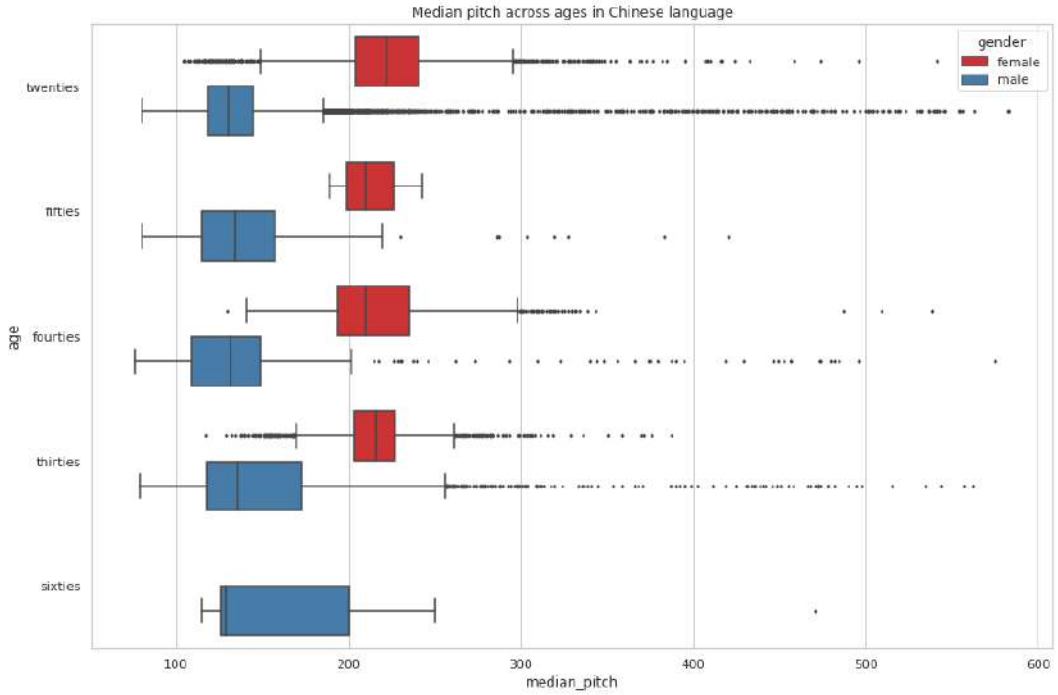


Figura B.5: La mediana del tono en los grupos de edad para el idioma Chino.

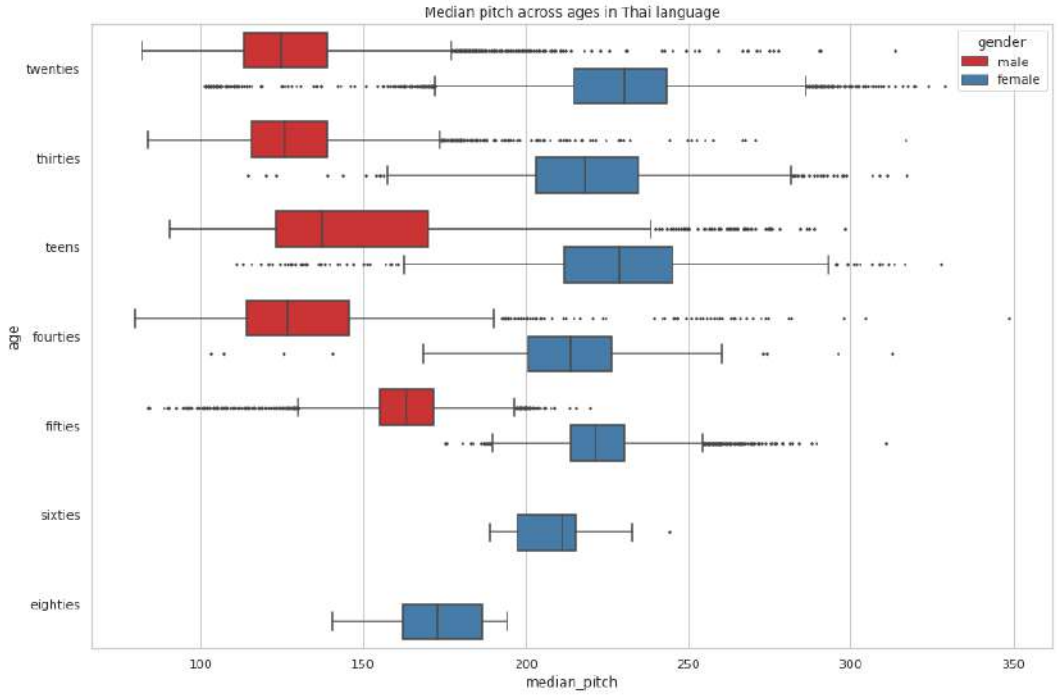


Figura B.6: La mediana del tono en los grupos de edad para el idioma Thai.

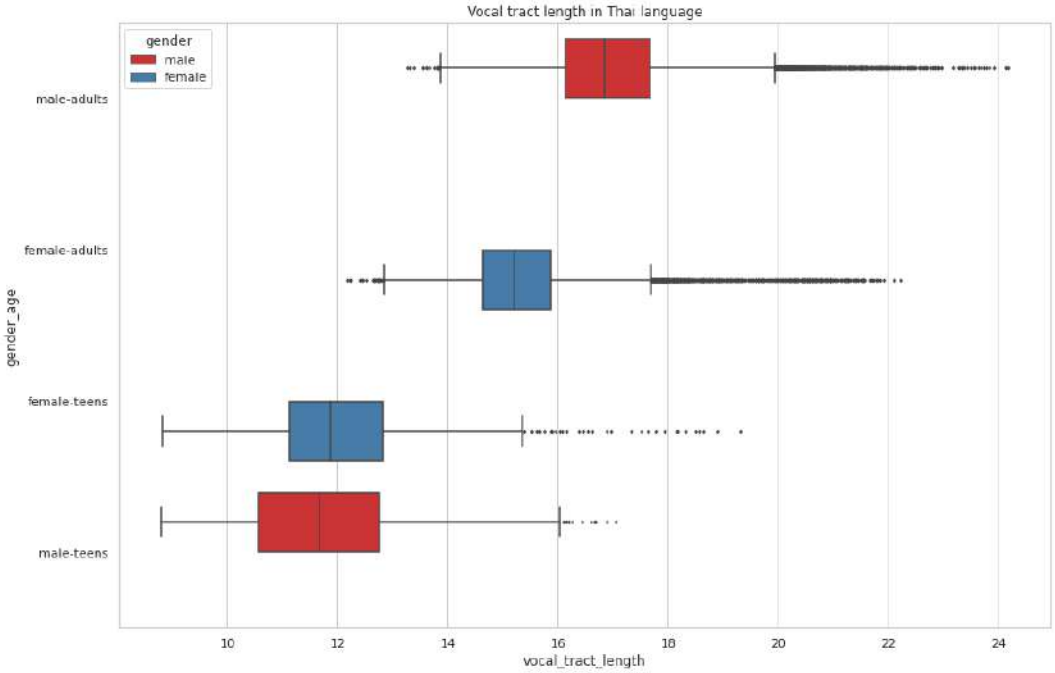


Figura B.7: Longitud del tracto vocal en los grupos de edad para el idioma Thai.

C | Algoritmo de Boersma para el cálculo del tono

En este trabajo se ha utilizado el lenguaje de programación Python para el procesamiento de la voz utilizando la librería Parselmount-Praat [56]. Esta librería utiliza los algoritmos implementados en el software PRAAT [53]. La extracción de las partes vocales de cada audio sigue el algoritmo de Boersman [53, 85] y el cálculo del tono de autocorrelación propuesto en [62].

Además, los parámetros utilizados aquí se basan en los utilizados en [55]. El procedimiento es el siguiente:

1. Los cuatro parámetros estándar se proporcionan para calcular una serie de candidatos de tono por ventana.
   a) Tiempo (*time step*) (t_s): Es el intervalo de medición (duración de la trama), en segundos. El valor estándar es 0, PRAAT analiza
   $$t_s = \frac{0{,}75}{P_f} \tag{C.1}$$
   muestras de audio por segundo. La variable P_f es el Pitch floor.
   b) Pitch floor (P_f): Los candidatos por debajo de esta frecuencia no serán reclutados. El valor estándar utilizado se fijó en $75Hz$.
   c) Longitud de la ventana (W_L): La longitud de la ventana analizada. Se calcula mediante
   $$W_L = \frac{3}{P_f}. \tag{C.2}$$
   Para este análisis, $W_L = 40$ milisegundos.
   d) Muy preciso: si está desactivado, la ventana es una ventana de Hanning con una longitud física de W_L. Si está activada, la ventana es una ventana gaussiana con una longitud física de $2W_L$. El valor utilizado fue off.
2. Un algoritmo de post-procesamiento busca el camino de menor costo a través de los candidatos de acuerdo con un funcional propuesto en [54]. Los parámetros que determinan dicho camino son:
   - Límite de tono (*Pitch ceiling*) P_c: Los candidatos por encima de esta frecuencia serán ignorados. El valor estándar utilizado fue $P_c = 350Hz$.
   - Umbral de silencio (*Silence threshold*) S_t: Los marcos que no contienen amplitudes por encima de este umbral (en relación con la amplitud máxima global), son probablemente silencio. El valor estándar utilizado fue $S_t = 0.03$.
   - Umbral de voz (*Voicing threshold*) V_t: La fuerza del candidato sin voz, en relación con la máxima autocorrelación posible. Si la cantidad de energía periódica en una trama es superior a esta fracción de la energía total (el resto es ruido), entonces Praat preferirá considerar esta marco como sonoro; en caso contrario, como no sonora. El valor estándar es $V_t = 0.45$.
   - Coste de la octava (*Octave cost*) O_c por octava: Grado de favorecimiento de los candidatos de alta frecuencia, en relación con la máxima autocorrelación posible. Esto es necesario porque, incluso en el caso de una señal perfectamente periódica, todos los subtonos de F_0 son candidatos igual de fuertes que la propia F_0. Para favorecer más el reclutamiento de candidatos de alta frecuencia, hay que aumentar este valor. El valor estándar es $O_c = 0.01$ por octava.
   - Coste del salto de octava (*Octave-jump cost*) O_j: Grado de desfavorecimiento de los cambios de tono, en relación con la máxima autocorrelación posible. El valor utilizado fue $O_j = 0.35$.
   - Coste sonoro/no sonoro (*Voiced/Unvoiced cost*) U_c: Grado de desfavorecimiento de las transiciones sonoras/no sonoras, en relación con la máxima autocorrelación posible. El valor utilizado fue $U_c = 0.14$.

3. Para cada intervalo sonoro detectado anteriormente, se encuentra un número de puntos de la siguiente manera:
   a) El primer punto t_1 es el extremo absoluto de la amplitud del archivo sonoro, entre $t_{mid} - \frac{T_0}{2}$ y $t_{mid} + \frac{T_0}{2}$, donde t_{mid} es el punto medio del intervalo y T_0 es el periodo en t_{mid}, como se puede interpolar a partir del contorno del tono.
   b) A partir de este punto, se busca recursivamente puntos t_i a la izquierda hasta alcanzar el borde izquierdo del intervalo. Estos puntos deben estar situados entre $[t_{i-1} - 1, 2T_0(t_{i-1}), t_{i-1} - 08T_0(t_{i-1})]$, y la correlación cruzada de la amplitud en su entorno $[t_i - \frac{T_0(ti)}{2}, t_i + \frac{T_0(ti)}{2}]$ con la amplitud del entorno del punto existente t_{i-1} debe ser máxima.
   c) Lo mismo se hace a la derecha de t_1.
   d) Se eliminan puntos si su valor de correlación es inferior a 0,3; además, se puede añadir un punto extra en el borde del intervalo de voz si su valor de correlación es superior a 0.7.