SISTEMA AUDIOVISUAL DE AYUDA PARA PERSONAS SORDAS

ANTHONY JONATHAN MUÑOZ CORREA

ESCUELA DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

FACULTAD DE INGENIERÍA

UNIVERSIDAD DEL VALLE

SANTIAGO DE CALI

2025

SISTEMA AUDIOVISUAL DE AYUDA PARA PERSONAS SORDAS

ANTHONY JONATHAN MUÑOZ CORREA

TRABAJO PRESENTADO PARA OPTAR TÍTULO DE INGENIERO ELECTRÓNICO

DIRECTOR

HUMBERTO LOAIZA CORREA, PH.D.

GRUPO DE INVESTIGACIÓN PERCEPCIÓN Y SISTEMAS INTELIGENTES

ESCUELA DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

FACULTAD DE INGENIERÍA

UNIVERSIDAD DEL VALLE

SANTIAGO DE CALI

2025

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

Nota de Aceptación

_____

_____

_____

Director

Jurado

Jurado

AGRADECIMIENTOS

Quiero expresar mi más profundo agradecimiento a mis padres y a mi hermana, cuyo apoyo incondicional y amor han sido un pilar fundamental en mi vida y formación académica; gracias a ellos soy quien soy hoy en día y en quien me quiero convertir. Agradezco de manera especial a mi director, Humberto Loaiza, por su guía, paciencia y compromiso durante todo el desarrollo de este proyecto. Finalmente, extiendo mi agradecimiento a la Universidad del Valle, mi alma mater, que me acogió desde el inicio de mi formación y me brindó las herramientas necesarias para crecer en la carrera. Con una mezcla de nostalgia al cerrar esta etapa académica, pero con la convicción de sobresalir profesionalmente. Llevaré siempre conmigo sus enseñanzas.

RESUMEN

Se presenta el desarrollo de un sistema audiovisual de ayuda para que personas sordas puedan ejercitar la pronunciación de palabras y suplir en cierta medida la retroalimentación perdida del oído por gráficas, indicadores de similitud y señales vibrotáctiles. Para ello, el sistema emplea reconocimiento automático de voz, calcula métricas de similitud de señales de audio y generación de señales vibrotáctiles. Este sistema está orientado al diseño de nuevas interfaces sensoriales para la percepción del lenguaje hablado mediante el tacto y la vista. El sistema permite identificar un conjunto de 12 palabras en español y traducirlas en patrones vibratorios únicos aplicables sobre la piel, con potencial aplicación en contextos de accesibilidad sensorial y rehabilitación del habla.

El módulo de reconocimiento de voz se basa en dos enfoques complementarios: la extracción de características acústicas predefinidas y el uso de modelos entrenados de representación profunda. En la primera estrategia, se extrajeron coeficientes cepstrales en las frecuencias de Mel (MFCC) junto con sus derivadas, espectrogramas y características chroma para construir vectores representativos, los cuales fueron utilizados en el entrenamiento máquinas de aprendizaje estadístico (HMM) como de aprendizaje profundo (LSTM). De forma paralela, se implementó un modelo preentrenado Wav2Vec2.0, capaz de transformar señales de audio crudas a 16 kHz en embeddings contextuales mediante una arquitectura de codificación convolucional seguida de bloques Transformer. Estos embeddings fueron posteriormente reducidos mediante análisis de componentes principales (PCA) para optimizar el uso de recursos computacionales y minimizar redundancias, siendo utilizados como entrada en máquinas de aprendizaje profundo (LSTM y GRU). Ambas estrategias fueron validadas utilizando un conjunto balanceado de 12 palabras, agrupadas según su punto de articulación (alveolares, bilabiales, palatales y velares), obteniéndose un F1-score promedio de 0.95 ± 0.02 con el enfoque de extracción de características predefinidas y un F1-score promedio de 0.85 ± 0.03 para el enfoque con Transformers.

La generación de señales vibrotáctiles se inspiró en el procesamiento auditivo humano, utilizando un banco de filtros Gammatone para descomponer cada señal en múltiples bandas de frecuencia relevantes para la percepción táctil. Sobre cada banda filtrada se aplicó la transformada de Hilbert para obtener la señal analítica, a partir de la cual se extrajo su envolvente temporal. Las envolventes de las distintas bandas fueron interpoladas, sumadas y normalizadas para formar una envolvente compuesta que representa la energía modulada del estímulo auditivo. Esta envolvente final fue utilizada para modular la amplitud de una portadora senoidal de 150 Hz, una frecuencia óptima para la estimulación táctil, generando así la señal. El Sistema de generación de señales vibrotáctiles se probó en seis personas logrando una precisión total promedio de 0.74 ± 0.07, destacando las palabras “*mucho*” y “*amigo*” con una precisión de 0.88 ± 0.07 y ± 0.06.

Palabras Clave: Procesamiento de señales, Reconocimiento de voz, Redes neuronales recurrentes, Señales vibrotáctiles, Estimulación multisensorial, interfaces para la percepción del lenguaje.

ABSTRACT

An audiovisual assistance system is presented to support deaf individuals in practising word pronunciation, partially compensating for the lost auditory feedback through visual graphs, similarity indicators, and vibrotactile signals. The system employs automatic speech recognition, calculates similarity metrics between audio signals, and generates vibrotactile stimuli. It is oriented towards the design of new sensory interfaces for perceiving spoken language through touch and sight. The system is capable of recognising a set of 12 Spanish words and translating them into unique vibratory patterns applicable to the skin, with potential applications in sensory accessibility and speech rehabilitation contexts.

The speech recognition module is based on two complementary approaches: the extraction of predefined acoustic features and the use of trained deep representation models. In the first approach, Mel-frequency cepstral coefficients (MFCCs), along with their derivatives, spectrograms, and chroma features, which were used to train both statistical learning models (HMM) and deep learning models (LSTM). In parallel, a pre-trained Wav2Vec2.0 model was implemented, capable of transforming raw 16 kHz audio signals into contextual embeddings through a convolutional encoder followed by Transformer blocks. These embeddings were subsequently reduced using Principal Component Analysis (PCA) to optimise computational resource usage and minimise redundancies, and were then used as input for deep learning models (LSTM and GRU). Both approaches were validated using a balanced dataset of 12 words, grouped according to their place of articulation (alveolar, bilabial, palatal, and velar), achieving an average F1-score of 0.95 ± 0.02 with the predefined feature extraction approach and average F1-score of 0.85 ± 0.03 for Transformers approach.

The generation of vibrotactile signals was inspired by human auditory processing, using a Gammatone filterbank to decompose each signal into multiple frequency bands relevant to tactile perception. For each filtered band, the Hilbert transform was applied to obtain the analytic signal, from which its temporal envelope was extracted. The envelopes of the different bands were interpolated, summed, and normalised to form a composite envelope representing the modulated energy of the auditory stimulus. This final envelope was used to modulate the amplitude of a 150 Hz sinusoidal carrier — an optimal frequency for tactile stimulation — thus generating the signal. The vibrotactile signal generation system was tested with six participants, achieving an average overall accuracy of 0.74, with the words *“mucho”* and *“amigo”* standing out with an accuracy of 0.88± 0.07 and ± 0.06.

Keywords: Signal processing, Speech recognition, Recurrent neural networks, Vibrotactile signals, Multisensory stimulation, Interfaces for language perception.

1. Introducción

La pérdida de audición, o hipoacusia, es uno de los problemas de salud crónicos más frecuentes a nivel mundial. Según la Organización Mundial de la Salud (OMS), más de 430 millones de personas presentan pérdida auditiva significativa, y se estima que para el año 2050 esta cifra superará los 750 millones de casos. Esta condición puede originarse por causas hereditarias, enfermedades, traumatismos, exposición prolongada al ruido o efectos secundarios de ciertos medicamentos. Si bien muchas de estas causas son prevenibles o tratables, su no atención oportuna tiene implicaciones directas en la calidad de vida de las personas, especialmente en lo relacionado con el desarrollo del lenguaje oral y la comunicación[1].

La imposibilidad de recibir retroalimentación auditiva durante la infancia limita la capacidad de las personas sordas para adquirir y perfeccionar la pronunciación. Si no se detecta y trata de manera temprana, esta situación puede comprometer profundamente su desarrollo comunicativo. Existen distintos tratamientos y dispositivos que buscan suplir o mitigar los efectos de la pérdida auditiva, como los audífonos, implantes cocleares o sistemas de amplificación sonora. No obstante, estas soluciones no siempre son accesibles o suficientes, especialmente en contextos de bajos recursos o cuando la pérdida auditiva es severa [2], [3].

En los últimos años, se han propuesto tecnologías que exploran rutas sensoriales alternativas a la audición, con el fin de proporcionar retroalimentación al usuario. Una de ellas es la estimulación vibrotáctil, que transforma señales sonoras en patrones de vibración perceptibles a través de la piel. Tecnologías como *Music: Not Impossible* permiten que personas con hipoacusia perciban la música mediante vibraciones localizadas en el cuerpo, como en chalecos, muñecas o tobillos [4]. De manera similar, en entornos terapéuticos, se han explorado dispositivos que permiten a los usuarios "sentir" el habla a través de la transformación de frecuencias sonoras en vibraciones táctiles, apoyadas por ayudas visuales y retroalimentación textual [5].

La problemática abordada se centra en la carencia de retroalimentación auditiva en personas sordas, la cual limita su desarrollo en la pronunciación del habla. Esta situación demanda sistemas alternativos que permitan sustituir, al menos parcialmente, la retroalimentación acústica mediante otros canales sensoriales, como el táctil, y que además proporcionen información visual clara y cuantificable sobre la precisión de la pronunciación.

Un sistema de apoyo a la rehabilitación del habla, basado en retroalimentación háptica, debe contemplar cuatro etapas fundamentales: adquisición y procesamiento de la señal de voz, reconocimiento automático del habla, análisis comparativo con grabaciones de referencia, y generación de salidas sensoriales, tanto visuales como táctiles, que proporcionen retroalimentación al usuario sobre la precisión de su pronunciación. Cada una de estas etapas implica desafíos técnicos que deben abordarse con metodologías adecuadas para garantizar un funcionamiento eficiente y robusto del sistema.

La etapa de adquisición y procesamiento requiere el uso de dispositivos capaces de

capturar con fidelidad las características acústicas de la voz, así como la implementación de técnicas de preprocesamiento que reduzcan el ruido y mejoren la calidad de la señal. La selección de filtros apropiados y algoritmos de normalización es crucial para evitar distorsiones que puedan afectar negativamente las siguientes fases del sistema.

El reconocimiento de voz es una etapa compleja en la que la selección de características acústicas juega un papel determinante. Estas características deben ser robustas frente a la variabilidad inherente de las grabaciones, como el timbre vocal, el ritmo de pronunciación o el ruido de fondo. Asimismo, el desempeño del clasificador depende de factores como el tipo de modelo de aprendizaje automático empleado, la calidad y cantidad de los datos de entrenamiento y la correcta configuración de hiperparámetros.

La tercera etapa corresponde al análisis comparativo entre la señal de voz del usuario y una grabación de referencia. Esta comparación debe ser capaz de captar diferencias fonéticas sutiles en la pronunciación, mediante el uso de métricas acústicas que reflejen de forma significativa el grado de similitud entre ambas señales.

Finalmente, la etapa de generación de salidas sensoriales plantea el reto de convertir la información acústica procesada en estímulos vibrotáctiles interpretables por el usuario. Esta conversión implica diseñar señales que conserven la información relevante de la palabra y que, al ser reproducidas mediante parlantes de baja frecuencia, puedan ser percibidas con claridad en zonas sensibles como la mano. La selección de las características vibrotáctiles adecuadas (frecuencia portadora, modulación, intensidad) condiciona la efectividad de esta retroalimentación.

Dadas estas dificultades, es necesario implementar métodos y algoritmos específicos en cada una de las etapas mencionadas, configurando los parámetros del sistema de manera cuidadosa para alcanzar un funcionamiento confiable. En este trabajo de grado se propone una solución basada en técnicas de procesamiento de señales, modelos de aprendizaje automático y desarrollo de software, orientada a la creación de un sistema de apoyo a la rehabilitación del habla en personas con discapacidad auditiva. Este trabajo busca responder la siguiente pregunta problematizadora:

¿Cómo generar un sistema que adquiera y procese señales de voz, realice comparaciones y genere indicadores de similitud con audios de referencias, y produzca vibraciones a través de un parlante, con potencial de sustituir parcialmente la retroalimentación auditiva y ser utilizado en procesos de rehabilitación del habla enfocándose en la pronunciación en personas sordas?

En este proyecto se realizará un desarrollo académico haciendo uso de dispositivos de uso común como un computador, micrófono y parlantes, en el cual, se probarán y evaluarán algunas técnicas de procesamiento para las etapas de un sistema de generación y comparación de sonido. Mediante el uso de bases de datos públicas o propias de sonido e imagen, y la colaboración de personas que no presenten afectaciones de habla y escucha, para configurar y realizar pruebas de cada etapa del

sistema. Debido a que con este trabajo de grado se pretende generar un primer desarrollo de ingeniería, no se hace necesario probar el sistema con personas sordas que estén en procesos de rehabilitación del habla.

1.1 Objetivos

1.1.1 Objetivo General

Desarrollar un sistema de comparación de señales de voz que genere indicadores visuales de similitud y vibraciones acústicas en un parlante como mecanismo de retroalimentación al usuario, utilizando técnicas de procesamiento de señales y reconocimiento de patrones, con potencial de ser utilizado en rehabilitación de personas sordas.

1.1.2 Objetivos Específicos

i. Documentar las palabras claves utilizadas en procesos de rehabilitación del habla en personas sordas y las características de vibraciones acústicas en parlantes que sean diferenciables al tacto del usuario.

ii. Implementar un sistema de reconocimiento de palabras en audios de voz adquiridos a través de un micrófono de computador.

iii. Obtener indicadores de similitud entre audios con palabras pronunciadas por un usuario y palabras de referencia empleadas en rehabilitación del habla en personas sordas.

iv. Establecer una interfaz intuitiva para visualización de indicadores de similitud y generación de vibraciones acústicas en parlantes diferenciables al tacto que sirvan de retroalimentación al usuario.

v. Evaluar el desempeño de cada una de las etapas desarrolladas y del sistema total.

1.2 Descripción del trabajo

Se implementó un sistema para apoyar procesos de rehabilitación del habla en personas con discapacidad auditiva, a través del análisis comparativo de señales de voz y la generación de retroalimentación háptica mediante vibraciones acústicas. El sistema fue diseñado bajo una arquitectura modular que contempla las etapas de adquisición, procesamiento de señales de voz, reconocimiento automático de palabras, visualización de señales, análisis de similitud acústica con grabaciones de referencia y, finalmente, la conversión de dicha información en señales vibrotáctiles reproducidas por parlantes de baja frecuencia.

Inicialmente, se realizó una revisión exhaustiva de la literatura, con el propósito de identificar los enfoques más efectivos en el reconocimiento de voz y la estimulación vibrotáctil como mecanismo complementario a la retroalimentación auditiva. Esta revisión permitió documentar el conjunto de palabras clave utilizadas en procesos de rehabilitación del habla, así como las características vibrotáctiles diferenciables al tacto según estudios previos. A partir de esta base teórica, se definieron los métodos de extracción de características, los modelos de aprendizaje automático y las estrategias de comparación acústica a implementar.

La señal de audio fue adquirida mediante un micrófono a una frecuencia de muestreo de 16 kHz en un solo canal (mono), en condiciones controladas para minimizar el ruido ambiental. Las grabaciones se normalizaron al rango de amplitud [-1, 1] para estandarizar la escala de procesamiento. Posteriormente, se aplicó un filtro pasa alta con una frecuencia de corte de 120 Hz con el fin de eliminar componentes de baja frecuencia no relevantes para el habla. Finalmente, se utilizó una técnica de reducción de ruido basada en sustracción espectral, que mejora la relación señal/ruido preservando las características acústicas esenciales de la voz.

El sistema fue desarrollado bajo dos enfoques principales para el reconocimiento de voz. En el primero, se extrajeron características acústicas predefinidas de las señales (MFCC y derivadas, espectrogramas y chroma), las cuales fueron normalizadas y utilizadas como entrada para dos clasificadores: Modelos Ocultos de Markov (HMM) y redes neuronales LSTM. En el segundo enfoque, se emplearon embeddings automáticos generados por el modelo preentrenado Wav2Vec2.0, los cuales encapsulan información fonética directamente desde el audio crudo. Estos vectores fueron clasificados mediante arquitecturas LSTM y GRU, optimizadas con búsqueda bayesiana de hiperparámetros a través de Optuna. En ambos casos, se evaluaron los modelos con métricas estándar como precisión, recall y F1-score, obteniendo valores superiores al 0.95 con la extracción de características predefinidas y rendimientos aceptables de 0.86 con embeddings automáticos.

La etapa de análisis comparativo de señales fue implementada mediante un cálculo de similitud acústica basado en distancias euclidianas ponderadas entre vectores normalizados. Estos indicadores se visualizan en la interfaz gráfica del sistema, junto con la representación de la señal y su clasificación. En paralelo, se diseñó un subsistema de estimulación vibrotáctil que convierte las señales de voz en pulsos táctiles mediante un banco de filtros Gammatone, transformada de Hilbert y modulación por amplitud con portadora de 150 Hz, permitiendo que el usuario perciba patrones temporales y espectrales asociados a cada palabra a través del sentido del tacto.

Finalmente, se integró una interfaz gráfica desarrollada en Python que permite grabar, almacenar, visualizar, clasificar y comparar las señales de voz, así como controlar la reproducción de vibraciones. El sistema fue validado experimentalmente con usuarios oyentes, alcanzando una tasa de reconocimiento promedio del 74.8%0.85 ± 7% en la identificación de 12 palabras por vía háptica, destacándose términos como “amigo” y “mucho” con más del 85% de aciertos.

Este desarrollo representa un primer prototipo funcional que demuestra la viabilidad de generar retroalimentación háptica a partir de señales de voz, como complemento en procesos de rehabilitación del habla, ofreciendo una alternativa accesible, no invasiva y escalable para personas con pérdida auditiva.

1.3 Organización de los capítulos

El documento consta de 5 capítulos cuyo contenido, se centra en aspectos particulares que contribuyeron al desarrollo del proyecto.

El capítulo 1 contiene la introducción general de los temas que abarca el documento, se exponen datos relevantes sobre las consecuencias de la pérdida de audición en las personas, precisamente en la disminución del habla, se habla de los métodos de clasificación y las dificultades en el análisis de estas. Además, se detalla el planteamiento del problema y se establece los objetivos generales y específicos del proyecto.

En el Capítulo 2 se registran los antecedentes nacionales e internacionales de proyectos similares, a partir de esto se incluye una discusión que permite identificar y seleccionar los métodos más adecuados, y con mejores resultados para cada etapa del proyecto. Después, se brinda la definición de todos los conceptos teóricos acerca del sistema y las técnicas que se implementan en los próximos capítulos.

En el capítulo 3 se plantea la solución propuesta para la implementación del sistema. Se detallan los procedimientos metodológicos necesarios para cumplir con los objetivos específicos del proyecto. El sistema propuesto es capaz de grabar una señal de voz y clasificarla entre un grupo de palabras seleccionadas, aplicando técnicas de filtrado para reducción de ruido y extracción de características predefinidas o a partir de modelos de representación profunda entrenados, el sistema permite la generación de señales vibrotáctiles asociadas a cada palabra, las cuales pueden ser reproducidas a través de un parlante. La interfaz gráfica facilita la interacción con el sistema, permitiendo al grabar y reproducir audios, visualizar imágenes y gráficas asociadas al reconocimiento, y gestionar la activación de las señales vibrotáctiles.

En el Capítulo 4 se describen las pruebas y se analizan los resultados de cada uno de los sistemas implementados. Luego se realiza un análisis comparativo del desempeño de los sistemas. Finalmente se establecen los alcances y las limitaciones del proyecto.

Una vez obtenido los resultados, en el *Capítulo 5* se definen las conclusiones generales del proyecto y se proponen perspectivas futuras y posibles ajustes que requiere el sistema para trabajos posteriores.

2. Antecedentes y Marco Teórico

En este capítulo se presentan los antecedentes regionales, nacionales einternacionales con resultados y metodologías relevantes para el desarrollo de este trabajo.

Las técnicas de procesamiento de señales y el proceso de reconocimiento de voz son fundamentales en el diseño de un sistema audiovisual destinado a la rehabilitación del habla en personas con discapacidad auditiva. Al procesar la señal se obtiene la extracción automática de características a partir de patrones sonoros y fonemas, permitiendo un análisis detallado de la pronunciación. La aplicación de algoritmos avanzados de reconocimiento de voz, basados en aprendizaje automático, entre los cuales se encuentran modelos como redes neuronales profundas, son empleados para entrenar modelos capaces de identificar matices y variaciones en el habla, que permiten descomponer la información auditiva y representar la señal de audio en un espacio de características más estructurado, facilitando la identificación de patrones de habla. Estos algoritmos mejoran la capacidad del sistema para generar mediciones y proporcionan retroalimentación por medio de vibraciones en la palma de la mano para cada usuario.

Finalmente, se presentan los conceptos necesarios sobre técnicas de procesamiento de señales, algoritmos, técnicas de aprendizaje automático para el sistema de reconocimiento de voz y la generación de señales sonoras capaces de generar vibraciones que puedan ser diferenciables por el tacto.

2.1 Antecedentes

En este apartado se presentan algunos de los antecedentes sobre técnicas de procesamiento de señales, detección de voz, comparación de sonidos y generación de vibraciones. La revisión permite identificar los elementos constitutivos y principales técnicas empleados en este tipo de proyectos. Se consultaron las bases de datos IEEE, Scopus, Science Direct, Scielo. Se realizó la búsqueda entre los años 2016 y 2024 diferenciando los trabajos nacionales e internacionales a excepción de un trabajo presentado en la universidad del Valle en 2011, utilizando las palabras clave: sordera, reconocimiento de voz, procesamiento, herramientas interactivas, realimentación aptica y señales vibrotactiles.

2.1.1 Antecedentes Nacionales

En [6] se implementó una interfaz de computador que interpreta gestos corporales mediante la captura de video mediante técnicas de visión artificial para el procesamiento y un sistema de reconocimiento de comandos usando modelos estocásticos. Para el reconocimiento de voz automático se implementó una red neuronal multicapa entrenada con los modelos ocultos de Markov (HMM).

El sistema fue probado en diferentes condiciones de ambientes controlados y semi controlados registrando resultados de desempeño para el reconocimiento de voz de 88,6677% y en la interpretación de gestos bajo condiciones favorables de iluminación

el desempeño en navegación fue del 93,604%, para el click del 96,970% y para la oclusión del 89,395%.

En [7] se presenta un método para identificar hablantes de manera semi-supervisada sin la necesidad de una etapa de entrenamiento previa. El método se basa en la extracción de características utilizando cocleagramas para palabras específicas seleccionadas. El sistema puede identificar uno o varios hablantes con una gran similitud a la prueba de audio o proporcionar una respuesta nula si ninguno de los hablantes cumple con un umbral de similitud. Los resultados del método propuesto se comparan con los obtenidos utilizando espectrogramas en lugar de cocleagramas. La evaluación del rendimiento se realiza mediante una matriz de confusión, y los resultados generales se expresan en términos de precisión general e índice kappa. Según las pruebas realizadas, el sistema propuesto muestra una precisión global superior al 97% y una índice kappa de alrededor de 0.78, indicando una alta confianza en los resultados de identificación y rechazo.

En [8] se utilizó API de código abierto para el reconocimiento de comandos de voz y se propone una arquitectura que permita la construcción de un sistema de reconocimiento de fonemas y sintetizadores de voz, además se utiliza normalización temporal la cual se basa en transformaciones lineales en el eje del tiempo para generar voces que puedan compararse con otras voces.

Para la comparación de patrones se utiliza la formación por clustering ya que ofrece soporte para varias versiones y diferentes ponentes y logrando una comparación directa entre la referencia y la señal de voz que se quiere reconocer. En el caso del generador de vectores característicos, utiliza los coeficientes cepstrales, lo que implica que se puede usar la distancia euclidiana en tales comparaciones [8]. Para las pruebas se obtuvieron grabaciones de 40 personas donde se obtuvo 27 interpretaciones correctas y 13 incorrectas obteniendo una eficiencia del 67,5%.

En [9] se implementó un sistema de comparación de palabras habladas y escritas utilizando técnicas de aprendizaje profundo, Se seleccionaron 10 palabras, las cuales se registraron en audio y se escribieron a mano frente a una cámara web. El objetivo era verificar si los datos de ambas modalidades coincidían y determinar si correspondían a la palabra requerida. Para lograr esto, se emplearon dos arquitecturas de redes neuronales convolucionales (CNN) diferentes, una diseñada para el reconocimiento de voz, que se basó en características obtenidas a partir de coeficientes cepstrales en las frecuencias de Mel (MFCC), y otra más rápida destinada a la escritura a mano, que localizaba e identificaba la palabra capturada.

Se llevaron a cabo pruebas en tiempo real, y los resultados mostraron una precisión general del 95.24%, lo que indica un buen rendimiento del sistema. Además, el sistema demostró ser rápido, con tiempos de respuesta de menos de 200 milisegundos durante la comparación.

En [10] se desarrolló un entorno virtual para un robot móvil capaz de reconocer comandos de voz básicos relacionados con la tarea de llevar o traer objetos en

entornos residenciales específicos. El sistema utiliza visión artificial para identificar los comandos de voz y los objetos con los que el robot debe interactuar, basándose en la captura de la escena en la que se encuentra el robot. Para la interacción voz-humano, se extraen los coeficientes cepstrales de Mel (MFCC) en conjunto con sus derivadas (deltas) y se utiliza una red neuronal convolucional de 6 capas convolucionales diseñada para reconocer los comandos relacionados con el transporte de objetos en un entorno residencial virtual. El algoritmo desarrollado logra un alto grado de éxito, con una tasa de acierto del 90%.

2.1.2 Antecedentes Internacionales

En [11] se evalúa y compara el rendimiento de cinco modelos de redes neuronales, red neuronal recurrente (RNN), memoria a corto plazo (LSTM) y unidad de puerta recurrente (GRU) en un sistema de reconocimiento de voz, para los modelos LSTM y GRU se toman dos caminos, unidireccional y bidireccional. Para el entrenamiento y test se usó la base de datos de comandos de voz proporcionado por tenserflow, el cual son ordenes simples pronunciadas por 20 usuarios 5 veces cada una, cada grabación fue Re muestreada a una frecuencia de 8Khz. Al evaluar los modelos con los datos de prueba se obtuvo una exactitud de 64.35% para el modelo RNN, 85.39% para LSTM, 89.96% para Bi-LSTM, 80.18% para GRU y 87.86% para Bi-GRU, aportando los mejores resultados el modelo LSTM bidireccional en el sistema de reconocimiento de voz.

En [12] se da a conocer una problemática que experimentan las personas con pérdida auditiva donde los subtítulos no logran expresar la emoción que proporciona la música en los medios audiovisuales. Para que las personas con pérdida auditiva puedan sentir una mejor experiencia se implementó la estimulación vibro táctil permitiendo transmitir la información emocional transmitida en los efectos de audio, encontrando una conexión entre el tacto y la audición, se conectan por los cuatro receptores táctiles existentes en la piel corpúsculos de Ruffini, Merkel, Meissner y Pacini, cada mecanismo cuenta con una sensibilidad a diferentes rangos de frecuencia permitiendo diferenciar entre tonos por medio del tacto.

En [13] Se desarrolla un sistema de reconocimiento de comandos de voz donde se usan redes neuronales convolucionales (CNN), las cuales han demostrado un gran rendimiento en esta área de aplicación. Se utiliza el conjunto de datos de comandos de voz que tiene un total de 105.829 expresiones en 35 clases, pronunciadas por 2618 hablantes. De esas 35 clases Se utilizan 38.546 comandos de voz en diez clases (sí, no, arriba, abajo, izquierda, derecha, en, apagado, parada, ir), con la proporción 80% para entrenamiento, 10% para prueba y 10% para validación. Para el análisis en frecuencia se utiliza la transformada discreta de Fourier (DFT) y obteniendo el espectrograma. Al realizar las pruebas del sistema se tiene un porcentaje de error del 3% y una precisión de clasificación del 97% para la red convolucional.

En [14] se propone un método para el reconocimiento de comandos de voz en modo de tiempo real utilizando una red neuronal profunda. que está construido con una combinación de diferentes arquitecturas de redes neuronales artificiales. Para la parte

de preprocesamiento de la señal se selecciona del oscilograma de una sola palabra, dividiéndola en fragmentos correspondientes a la pronunciación de las letras, obtención del espectro del fragmento de señal mediante la transformade de Fourier (DFT), selección de formantes de la señal de voz, y formación de una matriz de datos que describe la señal de voz correspondiente a una palabra de comando separada.

La obtención de un espectrograma que describe una sola palabra de un mensaje de voz permite usarlo para entrenar una red neuronal convolucional profunda. Al realizar evaluación de la eficiencia del reconocimiento se realiza mediante un análisis comparativo se obtuvieron los resultados precisión del reconocimiento de voz del 85%.

En [15] presenta un enfoque para clasificar voces normales y patológicas utilizando el modelo wav2vec 2.0 como extractor de características y comparando cuatro algoritmos de aprendizaje automático: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) y Decision Tree (DT). Utilizando la base de datos VOICED, con 208 grabaciones (150 patológicas y 58 normales) preprocesadas con técnicas de normalización y aumento de datos, wav2vec 2.0 generó representaciones contextuales de alta calidad a partir de señales de audio en bruto. Los clasificadores se entrenaron y evaluaron con validación cruzada estratificada, optimizando hiperparámetros mediante Grid Search. El modelo RF obtuvo el mejor rendimiento, con una precisión de 98%, recall de 97%, F1-score de 95% y AUC cercanas a 1.0, seguido del modelo DT, que destacó por su equilibrio entre precisión (97%) y recall (96%).

En [16] describe un sistema de reconocimiento de comandos de voz en tiempo real basado en la Distorsión Dinámica del Tiempo (DTW) y los Coeficientes Cepstrales de Frecuencia Mel (MFCC). Este sistema se enfoca en tareas de reconocimiento de comandos con un vocabulario específico. En tiempo real, procesa la voz mediante DTW, luego emplea una Red Neuronal Profunda (DNN) Para la extracción de características donde se convierte la señal original en el dominio del tiempo en una señal en el dominio de la frecuencia y realiza las transformaciones correspondientes para filtrar cierta información no relacionada con el contenido de la voz en la señal original para obtener el vector de características MFCC y calcular la distancia euclidiana entre los vectores de características de la voz de prueba y una plantilla. Los resultados experimentales indican que este sistema ofrece ventajas de rendimiento.

En el conjunto de datos usado se tomaron 200 palabras pronunciadas por diferentes personas divididas en género y edad para verificación, donde se obtuvo una precisión del 98% en todas las pruebas.

En [17] se presenta un modelo mejorado de red neuronal recurrente (RNN) con la arquitectura de memoria a corto plazo (LSTM) de aprendizaje profundo para abordar mejoras en los sistemas de reconocimiento de voz. En este nuevo enfoque, se incorpora una puerta de olvido (forget gate) en el bloque de memoria de la RNN, esto habilita al sistema para procesar flujos de entrada continuos de manera eficiente sin necesidad de aumentar significativamente los requisitos de ancho de banda. Los resultados mostraron que el modelo LSTM-RNN logro una precisión del 99.36% para el conjunto de datos de pruebas en inglés como referencia.

En [18] analiza los métodos y técnicas más efectivos para el aumento de datos en señales de audio. Las técnicas más recurrentes en el dominio del sonido incluyen la adición de ruido blanco o ambiental, el desplazamiento temporal (time shifting), y la modificación del tono (pitch), cada una orientada a incrementar la diversidad representacional sin alterar la semántica del contenido auditivo. Estas transformaciones permiten simular condiciones acústicas reales o variaciones naturales entre hablantes, contribuyendo así a la robustez y generalización de los modelos de aprendizaje profundo. En particular, la adición de ruido, presente en más del 39% de los estudios analizados, se consolida como una de las estrategias más efectivas para mitigar el sobreajuste y mejorar el desempeño en entornos ruidosos. El uso combinado de estas técnicas permite enriquecer los conjuntos de entrenamiento sin necesidad de recurrir a nuevas grabaciones, consolidándose como un componente esencial en el diseño de sistemas de reconocimiento y clasificación auditiva basados en redes neuronales.

En [19] presenta los patrones cutáneos de vibraciones en la interacción táctil con objetos, indicando que es posible decodificar los tipos de interacción directamente a partir de los patrones de vibración generados. Se notaron cambios rápidos en los patrones de intensidad a medida que las vibraciones se propagaban de manera desigual en la mano, la energía vibratoria alcanzó su punto máximo al contacto de un dedo con un objeto y luego se desvaneció rápidamente.

En [20] evalúa cómo la estimulación vibrotáctil basada en la envolvente temporal del habla puede mejorar la comprensión del lenguaje en entornos ruidosos. Utilizando una señal de 150 Hz modulada por la envolvente del habla extraída mediante un banco de filtros gammatone y transformada de Hilbert, los autores aplicaron esta estimulación en las palmas de las manos de 46 oyentes normales mientras escuchaban oraciones en francés con ruido de fondo. Se encontraron mejoras significativas en la discriminación del habla, particularmente en individuos con menor desempeño auditivo inicial y mayor sensibilidad táctil, aunque el efecto fue modesto (promedio de 0.41 dB de mejora). El estudio sugiere que esta forma de estimulación táctil, sencilla y de bajo costo computacional, puede apoyar la atención selectiva y percepción del habla, con potencial utilidad en poblaciones con dificultades auditivas o como complemento de implantes cocleares.

En [21] desarrollo un sistema de ayuda a la comunicación basado en dispositivos Android que convierte la voz en representaciones visuales e impulsos vibrotáctiles, y que a su vez permite que las personas sordas o mudas emitan mensajes mediante gestos o entradas táctiles que son transformadas en voz sintetizada. El sistema fue validado con seis participantes con discapacidad auditiva y dos oyentes, mostrando mejoras significativas tras un entrenamiento de dos horas: la precisión de la comunicación aumentó de un 45% a un 83% en tareas complejas, mientras que el tiempo de respuesta disminuyó notablemente. Asimismo, los usuarios calificaron positivamente la portabilidad, utilidad y facilidad de uso del dispositivo, con puntuaciones promedio superiores a 4.3 sobre 5, y expertos en el área destacaron su consistencia, usabilidad y adecuación al contexto real. Estos resultados demuestran la viabilidad de la retroalimentación vibrotáctil y visual como un canal complementario en

la comunicación.

En [22] diseñaron un sistema embebido implementado en un chaleco con motores de vibración, controlados mediante un microcontrolador Arduino UNO programado en Embedded C. El dispositivo emplea un micrófono para captar la señal acústica, la cual es procesada con la Transformada Rápida de Fourier (FFT) para identificar sus componentes frecuenciales y, posteriormente, convertirla en patrones de vibración distribuidos en diferentes zonas del chaleco. De esta manera, los usuarios pueden percibir variaciones sonoras a través del tacto, asociando la intensidad y frecuencia de los estímulos vibrotáctiles con características del sonido original. Los resultados experimentales evidenciaron la capacidad del sistema para transformar la voz y otras señales acústicas en impulsos vibratorios distinguibles, constituyendo una prueba funcional de la viabilidad de este enfoque. Los hallazgos respaldan el potencial de la retroalimentación vibrotáctil como un mecanismo práctico y accesible para apoyar la comunicación y la interacción cotidiana de personas sordas o mudas.

En [23] evaluaron el impacto de la estimulación vibrotáctil como alternativa sensorial para personas sordas o con hipoacusia severa, utilizando guantes hápticos sincronizados con estímulos audiovisuales y registrando la actividad cerebral mediante electroencefalografía (EEG). El estudio incluyó 35 participantes oyentes y 8 con pérdida auditiva, quienes visualizaron un cortometraje en tres condiciones: solo visión, visión con vibración y control audiovisual en oyentes. Los resultados mostraron que, bajo la condición de estimulación visual y vibrotáctil, las personas sordas presentaron activaciones corticales en las mismas áreas frontales y cinguladas que los oyentes con audio, pero con mayor intensidad estadísticamente significativa.

En [24] investigaron los efectos de la exposición a música vibrotáctil en personas con implante coclear, mediante un chaleco háptico con veinte actuadores sincronizados con la señal auditiva. El estudio incluyó veinticuatro participantes distribuidos en dos grupos, quienes fueron evaluados bajo dos condiciones: solo audio y audio acompañado de vibraciones. Los resultados mostraron mejoras estadísticamente significativas en las pruebas audiométricas de reconocimiento de tonos y palabras, tanto en silencio como en ruido, cuando se empleó la estimulación vibrotáctil en paralelo con el estímulo auditivo. Adicionalmente, la mayoría de los usuarios manifestó preferencia por la condición multimodal, destacando una experiencia más inmersiva, una mayor comprensión del ritmo y una mejor diferenciación instrumental.

En [25] evaluaron el efecto de la retroalimentación vibrotáctil congruente sobre la inteligibilidad del habla en condiciones de ruido, tanto en personas con audición normal como en usuarios de implante coclear. El experimento incluyó 22 oyentes normales y 14 usuarios de implante, quienes fueron expuestos a tres condiciones: solo audio, audio más vibración congruente (extraída de la envolvente temporal de la señal de habla) y audio más vibración incongruente (derivada de ruido modulado). Los resultados mostraron una mejora promedio de 5,3 % en el reconocimiento de palabras en ruido cuando la vibración era congruente con el habla, mientras que no se observaron beneficios en la condición incongruente, lo que confirma que el efecto no se debe a un estímulo táctil general, sino a la información temporal del habla transmitida

por la vibración.

En [26] exploraron el uso de retroalimentación vibrotáctil como apoyo a la localización de sonidos en personas sordas dentro de entornos de realidad virtual. Para ello, desarrollaron un traje háptico controlado por Arduino que incorporaba cuatro motores vibratorios distribuidos en distintas posiciones del cuerpo (muslos, brazos y orejas), los cuales se activaban de acuerdo con la dirección de una fuente sonora virtual. El estudio, realizado con veinte participantes sordos, demostró que el tiempo promedio para completar las tareas de localización disminuyó significativamente al emplear retroalimentación vibrotáctil en comparación con la condición sin traje, evidenciando una mejora en el desempeño de los usuarios. Asimismo, se observó que la colocación de los motores en brazos y orejas resultó más efectiva que en muslos, y los participantes expresaron una clara preferencia por estas configuraciones debido a la mayor naturalidad de la señal táctil.

2.1.3 Discusión

Los antecedentes mencionados presentan alternativas y enfoques en el desarrollo de sistemas de reconocimiento de voz. Cada sistema abarca diversas técnicas en las etapas esenciales del desarrollo del sistema además del enfoque de señales vibroáctiles percibidas a través de la palma de la mano. Se destaca el uso de bases de datos propias de los autores y datos recopilados por entidades abiertos al público. En el procesamiento de la señal se observa el uso de filtros para la reducción de ruido, aumento de datos mediante desplazamiento y velocidad de la señal. En cuanto a la extracción de características destacan técnicas como el análisis en frecuencia, que incluye la segmentación en ventanas y la aplicación de la transformada de Fourier (DFT) para obtener los coeficientes cepstrales de las frecuencias de Mel (MFCC) y generar espectrogramas. También se destaca el uso de máquinas de aprendizaje pre entrenados como wav2vec 2.0, que emplea un modelo basado en transformers para extraer y retornar características de alta calidad de señales. En la etapa de clasificación se observa el enfoque en sistemas de máquinas superficiales y profundas; Entre los sistemas superficiales se encuentra los modelos ocultos de Markov (HMM). Por otro lado, en las máquinas de aprendizaje profundo, el uso de redes neuronales recurrentes (RNN) y sus variantes como LSTM y GRU han demostrado un desempeño notable en la tarea del reconocimiento de voz, alcanzando precisiones superiores al 80%. En la etapa de generación de vibraciones acústicas, los patrones cutáneos generados durante la interacción táctil están controlados por los corpúsculos presentes en la piel de la mano, los cuales presentan mayor susceptibilidad en frecuencias entre los 5 a 400 Hz. La retroalimentación vibrotáctil es un canal sensorial útil y viable, capaz de mejorar la percepción del habla en ruido, apoyar la localización de sonidos y complementar la experiencia auditiva en personas con discapacidad. Se ha demostrado que transforma señales acústicas en vibraciones comprensibles, activas regiones cerebrales asociadas a la atención y facilita la comunicación en diferentes contextos

En la Tabla 2.1, se resumen los aspectos más relevantes de algunos de los antecedentes consultados que sirvieron de soporte para el desarrollo de este trabajo.

Tabla 2.1 Comparativa de antecedentes consultados para sistemas de reconocimiento de voz.

Fuente: Elaboración propia.

Referencia	Base de datos	Técnica de procesamiento	Características	Clasificador	Dispositivo de adquisición	Desempeño	Métrica de desempeño
[6]	20 palabras	DTFT	MFCC (13 coeficientes)	HMM	Interfaz	88,6677%	Acierto
[8]	40 palabras	DTFT	MFCC (10 coeficientes)	GRU y LSTM	Interfaz	67.5%	Eficiencia
[9]	10 palabras	DTFT	MFCC (13 coeficientes)	CNN	Micrófono	95,24%	Precisión
[10]	50 palabras	DTFT	MFCC y Deltas	CNN	Interfaz	90%	Acierto
[11]	20 palabras	DTFT	MFCC y Deltas	LSTM	-	89.96%	Acierto
[13]	10 palabras	DTFT	Espectrograma	CNN	-	97%	Precisión
[14]	25 palabras	DTFT	Espectrograma	DNN	Interfaz	85%	Precisión
[15]	208 palabras	Wav2vec 2.0	Embeddings	RF	-	97%	Recall
[16]	200 palabras	DTW	MFCC	DTW	Micrófono	98%	Precisión
[17]	240 palabras	DFTT	MFCC	LSTM	Microfono	99.36%	Precisión

2.2 Marco Teórico

El objetivo general de este trabajo es detectar automáticamente una palabra en una señal de voz y obtener su semejante visual para generar una señal de audio capaz de generar vibraciones acústicas mediante técnicas de procesamiento de señales y máquinas de aprendizaje. Para realizar este trabajo es necesario tener en cuenta algunos conceptos generales sobre la sordera y los métodos de rehabilitación, el procesamiento de señales, la clasificación automática por aprendizaje automático y la generación de señales que generen vibraciones susceptibles al tacto.

2.2.1 Conceptos generales

2.2.1.1 Sordera

La pérdida auditiva ocurre cuando una parte del oído o del sistema auditivo no funciona de la manera habitual. En la Figura 2.1 se ilustran las principales estructuras del oído, divididas en oído externo, medio e interno. Existen cuatro tipos de pérdida auditiva: conductiva, neurosensorial, mixta, trastorno del espectro de la neuropatía auditiva y a diferentes grados [27].

- Pérdida auditiva conductiva: Causada por algo que impide que los sonidos pasen por el oído externo o medio, se puede tratar con medicamentos o cirugía.

- Pérdida auditiva neurosensorial: Ocurre cuando hay un problema en la forma en que funciona el oído interno o el nervio auditivo.

- Pérdida auditiva mixta: Incluye la pérdida auditiva tanto conductiva como neurosensorial.

- Trastorno del espectro de la neuropatía auditiva: se produce cuando el sonido entra normalmente en el oído, pero debido a un daño en el oído interno o en el nervio auditivo, el sonido no está organizado de forma que el cerebro pueda comprenderlo.

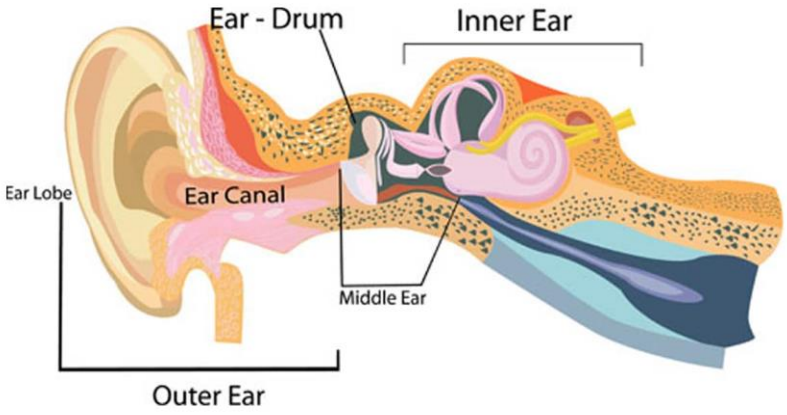

Figura 2.1 Partes del oído.

Fuente: [27].

Los métodos de tratamiento varían del momento en el que la persona es diagnosticada con pérdida de audición y el nivel que esté presente, entre más rápido sea el diagnóstico y se inicie la rehabilitación el desarrollo cognitivo y lingüístico será exitoso

2.2.1.2 Test de articulación a la repetición (T.A.R)

El test de articulación a la repetición es una prueba que permite evaluar el nivel fonético de personas las cuales no han desarrollado el habla por cualquier trastorno o discapacidad, logrando detectar posibles inconsistencias fonémicas y evaluar la memoria auditiva [28].

La prueba aborda fonemas categorizados según su punto articulatorio y métrica en palabras, presentados de acuerdo con la posición silábica en palabras con significado como se observa en la Tabla 2.2. Además, incorpora la evaluación de difonos vocálicos, consonánticos, palabras polisilábicas y frases con metría ascendente. Esta

prueba es adecuada para ser administrado a cualquier persona que haya alcanzado el desarrollo del lenguaje y tenga la capacidad de repetir palabras; además, no presenta restricciones de edad en su aplicación [28].

Tabla 2.2 Palabras usadas en el test de articulación a la repetición

Fuente: [28].

Fonemas		**Silaba inicial**	**Silaba Media**	**Silaba Final**	**Silaba Trabada**
Bilabiales	B	Bote	Cabeza	Nube	Objeto
		Bala	Tabaco	Tubo	Submarino
	P	Pato	Zapato	Copa	Apto
		Peso	Tapado	Sopa	Séptimo
	M	Mano	Camisa	Suma	Campo
		Mucho	Camote	Lomo	Temprano
Labio dental	F	Foca	Búfalo	Café	Aftosa
		Fino	Zafiro	Mofa	Difteria
Dentales	D	Dama	Cadena	Codo	Pared
		Dato	Madera	Nudo	Admite
	T	Tapa	Botella	Mata	Etna
		Tina	Tetera	Lote	-
Alveolares	S	Sol	Cocina	Casa	Pasto
	N	Nido	Panera	Maní	Canto
	L	Luna	Caluga	Pala	Dulce
	R	-	Para	Coro	Torta
	RR	Rosa	Carrosa	Perro	-
Palatales	Y	Lluvia	Payaso	Calle	-
	Ñ	Ñato	Puñete	Caña	-
	CH	Chala	Lechuga	Noche	-
Velares	K	Casa	Paquete	Taco	Acto

	G	Gato	Amigo	Jugo	Signo
	J	José	Tarjeta	Caja	Reloj
Difonos Vocálicos					
Piano			Violín	Diuca	
Vaina			Auto	Fui	
Difonos Consonánticos					
Tabla	Clavo		Flecha	Dragon	
Regla	Braza		Fruta	Crema	
Premio	Atlas		Tigre	Plato	
Palabras Polisilábicas					
Carabinero			Ametralladora		
Panadería			Helicóptero		
Caperucita			Bicicleta		

2.2.1.3 Método verbo tonal

Este método toma elementos de la percepción auditiva y táctil en donde la palabra hablada se usa como estímulo, se ejecuta por medio de aparatos SUVAG (Sistema Universal Verbal Auditivo de Guberina), los cuales son amplificadores de sonidos con vibradores, micrófonos y auriculares. Un principio básico de la investigación y rehabilitación verbo tonal es que todo el cuerpo humano funciona como receptor y transmisor, el cuerpo humano vibra con frecuencias entre 7 y 400Hz, por esto el procedimiento básico inicial en la rehabilitación de la persona con déficit auditivo se apoya en la vibración [5].

La piel humana contiene fibras de mielinizadas que responden a los estímulos mecánicos y la intensidad de los estímulos se correlaciona con su frecuencia de descarga. Estas fibras terminan en corpúsculos de Merkel, Meissner, Pacini o Ruffini [29].

Los corpúsculos Merkel-SA1 (tipo I de adaptación lenta) en el dedo son sensibles a la fuerza constante, a la baja frecuencia (f < 5 Hz), Tienen una mayor sensibilidad a las características y curvaturas de la superficie [29].

Los corpúsculos de Meissner-FAI (Fast Adapting type I) son cuatro veces más sensibles a la deformación/movimiento dinámico de la piel que los corpúsculos SA1. Pueden detectar fuerzas repentinas asociadas con objetos de mano y responden a

presiones y vibraciones entre 5 y 50 Hz [29].

Los corpúsculos de Ruffini-SAII (Adaptación Lenta tipo II) proporcionan información sobre la dirección del movimiento o la fuerza, particularmente cuando el movimiento implica el estiramiento de la piel. Son sensibles al estiramiento de la piel y a las fuerzas constantes. Los corpúsculos de Pacinian-FAII (tipo II de adaptación rápida) están hechos para capturar grandes tensiones y tensiones de baja frecuencia que se encuentran en las actividades manuales diarias. Tienen una resolución espacial muy baja y responden a estímulos distantes. Los FAII son sensibles a las deformaciones micrométricas y a los estímulos vibro táctiles en el rango de 40 Hz a > 400 [29].

2.2.2 Técnicas de procesamiento

El primer paso en la mayoría de las aplicaciones de procesamiento de voz digital es convertir la forma de onda acústica en una secuencia de números. La mayoría de los convertidores analógico a digital modernos funcionan mediante el muestreo a una tasa muy alta, aplicando un filtro digital pasa bajos con un punto de corte establecido para preservar un ancho de banda prescrito, y luego reduciendo la tasa de muestreo a la tasa de muestreo deseada [30].

2.2.2.1 Transformada de Fourier de tiempo Discreto (DTFT)

La Transformada de Fourier es una herramienta matemática utilizada para analizar señales y funciones en el dominio de la frecuencia. Convierte una señal en el dominio del tiempo en su representación en el dominio de la frecuencia. Esto permite analizar la contribución de cada componente frecuencial en una señal y estudiar sus propiedades [31]. Es un importante método en el campo de las mediciones de audio y la acústica y se calcula a partir de la ecuación 2.1.

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n) e^{-i\omega n} \qquad \textit{Ecuación 2.1}$$

Donde: $x(n)$ es la señal de entrada

$X(\omega)$ es la transformada de Fourier

n es el índice que recorre los posibles valores de la sumatoria

2.2.2.2 Filtro pasa altas y Butterworth

En procesamiento de señales un filtro pasa altas permite el paso de frecuencias altas mientras que atenúa o elimina las frecuencias bajas por debajo de un umbral conocido como frecuencia de corte. Las frecuencias no se eliminan en un instante; en lugar de eso, hay una transición suave en torno a la frecuencia de corte conocida como pendiente o roll-off que se mide en decibeles (dB). Los filtros pasa altas son esenciales para eliminar ruidos de fondo de baja frecuencia y destacar señales útiles que se encuentren en rangos más altos, como la vocalización humana[32].

El filtro Butterworth es un tipo de diseñado para presentar una respuesta de magnitud

lo más plana posible en la banda pasante, sin ondulaciones ni picos, y una atenuación progresiva en la banda de rechazo. Esta característica lo hace especialmente adecuado para aplicaciones de procesamiento de señales sensibles, como el habla, donde es fundamental preservar la forma espectral sin introducir distorsiones. Su comportamiento está determinado por el orden del filtro (n), que corresponde al número de polos del sistema y define la pendiente de atenuación en la banda de rechazo, la cual aumenta aproximadamente 20 dB por década por cada orden. Así, a mayor orden, el filtro ofrece una transición más abrupta entre la banda pasante y la banda de rechazo, permitiendo eliminar de manera más eficaz componentes no deseadas cercanas a la frecuencia de corte, aunque a costa de un mayor costo computacional y complejidad de implementación [33].

2.2.2.3 Reducción de ruido a través de espectrograma

La reducción de ruido en señales de audio permite analizar cómo se distribuyen las frecuencias de una señal a lo largo del tiempo, facilitando la identificación y eliminación de componentes no deseados.

La Transformada de Fourier de Tiempo Corto (STFT) es una técnica de procesamiento de señales para la reducción de ruido, ya que permite analizar la distribución de frecuencias de una señal en función del tiempo mediante un espectrograma, facilita la identificación y atenuación de componentes no deseadas, como ruido constante o picos en bandas específicas, mediante filtrado selectivo en el dominio tiempo-frecuencia. Se puede usar para la eliminación de ruidos constantes como zumbidos, filtrado de alta frecuencia y sustracción espectral. Posteriormente, la señal se reconstruye en el dominio temporal usando la Transformada de Fourier Inversa (ISTFT), permitiendo un análisis detallado, ideal para señales no estacionarias como el audio [34].

2.2.2.4 Ventana Hanning

La ventana de Hanning es una herramienta matemática que se utiliza en el procesamiento de señales para limitar la duración de una señal y reducir las fugas espectrales al aplicar transformadas de Fourier en señales finitas. Su forma de campana se define matemáticamente y se caracteriza por atenuar los lóbulos laterales en la transformada de Fourier, lo que ayuda a centrar la energía en la frecuencia principal de la señal. Aunque tiene una respuesta en frecuencia más estrecha en comparación con la ventana rectangular, mantiene una resolución aceptable en frecuencia. La ventana de Hanning es comúnmente utilizada en aplicaciones de análisis espectral y procesamiento de señales donde se busca un equilibrio entre la atenuación de las fugas espectrales y la resolución en frecuencia [35].

2.2.2.5 Normalización Z-Score

La normalización Z-Score es una técnica de escalado estadístico que transforma los datos para que cada característica tenga una media de cero y una desviación estándar de uno. Se calcula restando a cada valor la media de la característica y dividiendo el resultado entre su desviación estándar. Este procedimiento permite que las variables

con diferentes escalas o unidades sean comparables entre sí, evitando que aquellas con valores de mayor magnitud dominen el proceso de aprendizaje de los modelos. Además, la normalización Z-Score mejora la estabilidad numérica y la velocidad de convergencia de los algoritmos de optimización, por lo que es ampliamente utilizada en tareas de procesamiento de señales y aprendizaje automático [36].

2.2.3 Aprendizaje automático (ML)

El aprendizaje automático (ML) se centra en el desarrollo de algoritmos y modelos que permiten el aprendizaje de patrones para realizar tareas específicas y la toma de decisiones o predicciones basadas en los datos de entrada. Los algoritmos se entrenan utilizando conjuntos de datos, que consisten en ejemplos de entrada y las respuestas o salidas esperadas, durante el entrenamiento el modelo ajusta sus parámetros para minimizar la diferencia entre las predicciones y las respuestas reales [37].

Este proceso se lleva a cabo en dos etapas, donde un sistema evalúa y coteja las características pertinentes de un objeto o evento con aquellas ya conocidas. Si se detectan discrepancias significativas, el sistema ajusta su modelo del objeto o evento. Este proceso de aprendizaje automático contribuye a mejorar el rendimiento del sistema y puede emplear diversas técnicas matemáticas y búsquedas en bases de datos. A diferencia de los procesos cognitivos humanos, el aprendizaje automático no está restringido a ellos y requiere estructuras de representación del conocimiento apropiadas para identificar hechos relevantes [38].

Para determinar el algoritmo de clasificación se requiere conocer las características de clasificación de las señales y definir si se requiere aprendizaje supervisado, no supervisado, semi supervisado.

El aprendizaje superficial se basa en el aprendizaje automático donde se aprende por los datos con descripción de características ya definidas. Para la selección y extracción de características se requiere recopilar conocimientos con el fin de extraer características de los datos originales, generando una mayor complejidad computacional [39].

El aprendizaje profundo está relacionado con la inteligencia artificial, basado en redes neuronales profundas para el aprendizaje de patrones y obtener características del conjunto de datos proporcionado. El aprendizaje profundo aborda problemas complejos en el campo de reconocimiento de imágenes, procesamiento de lenguaje y habla [40].

2.2.3.1 Reconocimiento automático de voz (ASR)

El reconocimiento de la voz consiste en tomar como entrada la forma acústica de la voz humana y producir una salida u orden equivalente, para lograr dicho resultado, la señal de voz ingresa a un módulo de procesamiento de señales donde se toma la señal de audio analógica y se extraen los vectores de características sobresalientes y se comparen con una base de datos de vocabulario, palabras y sílabas, donde se reconozcan los patrones de habla [41].

En la actualidad el reconocimiento de voz ha tomado mucho auge en el desarrollo de tecnologías para seguridad, interpretación del lenguaje para transcripciones automáticas, asistentes virtuales como Siri, Cortana. Avanzando significativamente en las últimas décadas gracias al desarrollo de técnicas más sofisticadas de procesamiento de señales, así como al uso de modelos de aprendizaje profundo, lo que permite la mejora en precisión y la capacidad de adaptación de los sistemas de reconocimiento de voz, permitiendo una interacción más natural entre los usuarios y las máquinas.

2.2.3.2 Modelos ocultos de Márkov (HMM)

Los modelos ocultos de Márkov (HMM) describen modelos estadísticos con transiciones de probabilidad y se usa para describir sistemas que cambian a lo largo del tiempo, donde el estado actual no se puede observar directamente, sino a través de símbolos o eventos que dependen del estado. Estos modelos resultan beneficiosos para examinar una secuencia de observaciones en intervalos discretos, como una secuencia de muestras acústicas obtenidas de una señal de voz. Estos modelos son muy apropiados para el reconocimiento de voz debido a su capacidad inherente de representar eventos acústicos de duración variable y a la existencia de algoritmos eficaces para computar automáticamente los parámetros del modelo a partir de los datos de entrenamiento [42].

La matriz de probabilidad está dada por la ecuación 2.2, donde a_{ij} es la probabilidad de tomar una transición desde el estado i hasta el estado j.

$$a_{ij} = P\left(s_t = j \middle| s_t = i\right) \qquad \textit{Ecuación 2.2}$$

La matriz de probabilidad de emisión está dada por la ecuación 2.3, donde b_{ik} es la probabilidad de emitir un símbolo cuando se entra en un estado i .

$$B = \{b_i(k)\} \qquad \textit{Ecuación 2.3}$$

2.2.3.2.1 Algoritmo Baum-Welch

El algoritmo Baum–Welch es un método de entrenamiento utilizado para estimar los parámetros de los Modelos Ocultos de Márkov (HMM) cuando las secuencias de estados ocultos no son observables directamente. Se basa en el algoritmo de Expectation-Maximization (EM) y realiza un proceso iterativo de dos fases: en la fase de expectación calcula las probabilidades esperadas de transición y emisión a partir de los datos observados, y en la fase de maximización ajusta los parámetros del modelo (matriz de transición, probabilidades iniciales y distribuciones de emisión) para maximizar la verosimilitud de las secuencias de entrenamiento [43].

2.2.3.3 Coeficientes cepstrales de las frecuencias de Mel (MFCC) y derivada

En la actualidad, la mayoría de los sistemas de reconocimiento automático del habla (ASR) utilizan los coeficientes cepstrales de frecuencia Mel (MFCC). En MFCC, el

espectro de voz preprocesado se pasa a través de un banco de filtros Mel, que es un banco de filtros triangular. La energía de salida se comprime logarítmicamente y se transforma al dominio cepstral mediante una transformación coseno discreta inversa. Los MFCC son coeficientes para la representación del habla basados en la percepción auditiva humana. Estos surgen de la necesidad, en el área del reconocimiento de audio automático, de extraer características de las componentes de una señal de audio que sean adecuadas para la identificación de contenido relevante, así como obviar todas aquellas que posean información poco valiosa como el ruido de fondo, emociones, volumen, tono, etc. y que no aportan nada al proceso de reconocimiento. Se representa en la ecuación 2.4 [44].

$$Mel(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right) \qquad \textit{Ecuación 2.4}$$

En la ecuación, f se refiere a la frecuencia normal y Mel (f) es la frecuencia de salida de la escala Mel. Por lo general, esta escala cubre el rango de frecuencia de 156 Hz a 6844 Hz y es logarítmica por encima de 1 KHz y lineal por debajo del rango de frecuencia de 1 KHz. En general, los primeros 13 coeficientes son suficientemente amplios para representar la señal de voz [44].

Para enriquecer esta representación y permitir que los modelos de reconocimiento capturen los patrones de cambio a lo largo del tiempo, se introducen las derivadas de primer y segundo orden de los MFCC, comúnmente denominadas deltas y delta-deltas.

Las deltas corresponden a una estimación de la derivada temporal de los MFCC, y su objetivo es reflejar la velocidad con la que cambian dichos coeficientes en el tiempo. A su vez, las delta-deltas representan la derivada de segundo orden, es decir, la aceleración del cambio de los coeficientes. Estas derivadas se calculan mediante una media ponderada de las diferencias entre coeficientes MFCC en una ventana de tiempo centrada en el instante actual [45]. La fórmula utilizada para calcular las derivas se presenta en la ecuación 2.5 donde N define el tamaño de la ventana en ambos sentidos temporales.

$$\Delta_{ct} = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \qquad \textit{Ecuación 2.5}$$

2.2.3.4 Espectrograma

El espectrograma consiste en la representación gráfica del espectro de frecuencias de la emisión sonora. En la Figura 2.2. se ilustra el espectrograma donde revela rasgos, como altas frecuencias o modulaciones de amplitud, que no pueden apreciarse incluso aunque estén dentro de los límites de frecuencia del oído humano. Normalmente, un espectrograma representa el tiempo sobre el eje horizontal, la frecuencia sobre el eje vertical y la amplitud de las señales mediante una escala de colores [46].

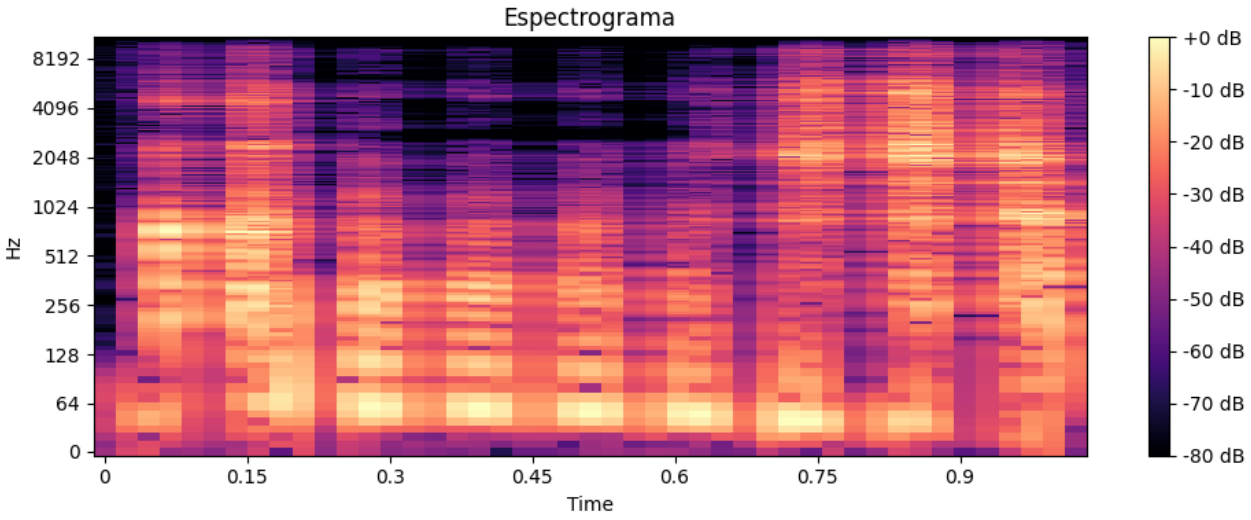

Figura 2.2 Espectrograma de una señal de audio.

Fuente: Elaboración propia.

2.2.3.5 Chroma

El chroma es una calidad asociada de las clases de tono, puede descomponerse en un valor independiente denominado chroma y en altura tonal, la cual indica en que octava se encuentra el tono. Un vector chroma es un vector de 12 elementos que representa la cantidad de energía presente en cada clase de tono en una señal. Este vector es una representación basada en la percepción del tono musical [47].

2.2.3.5.1 Características chroma

El termino características chroma está relacionado con las 12 clases de tonos diferentes. Una propiedad principal es la representación del contenido armónico y melódico de una ventana de tiempo corta del audio. Estas características se extraen del espectro de magnitud utilizando transformada de Fourier de tiempo corto (STFT). En el proceso de extracción de características chroma, toda la información espectral asociada a una clase de tono específica se agrupa en un único coeficiente. Este enfoque se emplea con frecuencia en tareas como análisis estructural del audio. Además, las características chroma han demostrado ser una representación de nivel intermedio altamente eficaz para la recuperación de audio basada en contenido, como la coincidencia de audio [47].

2.2.3.6 Modelo wav2vec 2.0

El modelo wav2vec 2.0 es una arquitectura basada en redes neuronales profundas, diseñada para extraer representaciones ricas y robustas de señales de voz. Procesa directamente el audio en forma de ondas, en su núcleo el modelo utiliza una combinación de una red convolucional inicial para capturar patrones locales de la señal y una red Transformers que modela relaciones a largo alcance en las señales temporales. Durante su entrenamiento, se emplea una técnica de aprendizaje

contrastivo en la que partes de la señal son enmascaradas, y el modelo aprende a predecir las representaciones de las porciones enmascaradas utilizando contexto no enmascarado, optimizando así su capacidad de aprendizaje contextual. Las representaciones generadas por Wav2Vec son embeddings de alta dimensionalidad que encapsulan tanto la información acústica como fonética [15].

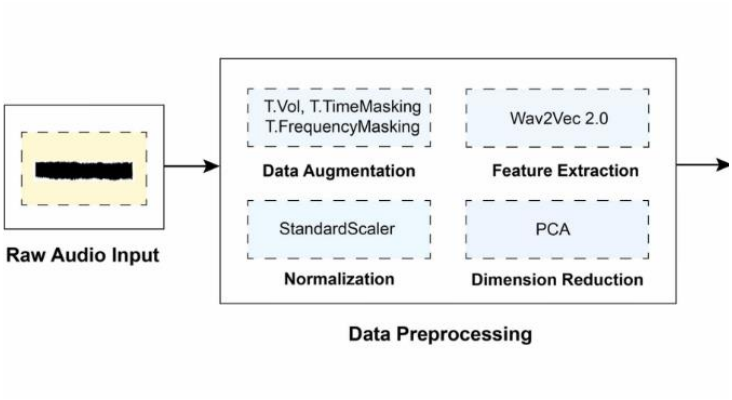

Figura 2.3 Estructura del modelo wav2vec para la extracción de características.

Fuente: [15].

2.2.3.7 Redes neuronales

Las redes neuronales se componen de nodos o neuronas que están conectados entre sí, cada nodo recibe una cantidad específica de entradas y produce una salida. En las capas de salida, cada nodo realiza un cálculo de suma ponderada con los valores recibidos de los nodos de entrada y se generan salidas utilizando funciones de transformación no lineales aplicadas a estas sumas. Las redes neuronales se inspiran en la forma en que las neuronas biológicas se comunican entre sí y se adaptan al aprendizaje. Las redes neuronales se utilizan para resolver problemas complejos que requieren inteligencia artificial, como el reconocimiento de imágenes, el procesamiento de lenguaje natural, la predicción de datos o el control de sistemas dinámicos [48].

2.2.3.8 Redes neuronales recurrentes y de memoria a corto plazo

Dado que gran parte del procesamiento del lenguaje natural (PNL) depende del orden de las palabras u otros elementos como fonemas u oraciones, es útil tener memoria de los elementos anteriores al procesar los nuevos. A veces existen dependencias hacia atrás, es decir, el procesamiento correcto de algunas palabras puede depender de las palabras que siguen. Por lo tanto, es beneficioso observar oraciones en ambas direcciones, hacia adelante y hacia atrás, utilizando dos capas RNN y combinando sus salidas. En las redes de memoria corto plazo (LSTM) Figura 2.4, los nodos recursivos están compuestos por varias neuronas individuales conectadas de una manera diseñada para retener, olvidar o exponer información específica. Mientras que los RNN genéricos con neuronas individuales que se retroalimentan a sí mismas técnicamente tienen cierta memoria de resultados pasados hace mucho tiempo, estos resultados se diluyen con cada iteración sucesiva. A menudo, es importante recordar información del

pasado lejano, mientras que, al mismo tiempo, otra información muy reciente puede no ser importante. Al utilizar bloques LSTM, esta información importante se puede retener por mucho más tiempo, mientras que la información irrelevante se puede olvidar [48].

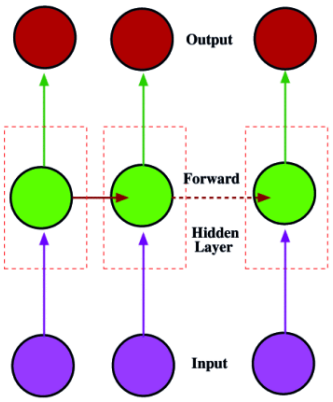

Figura 2.4 Estructura básica del modelo LSTM.

Fuente: [11].

Las GRU son un tipo de red neuronal recurrente que utiliza mecanismos de compuerta para controlar el flujo de información y manejar dependencias a largo plazo en secuencias de datos. En la Figura 2.5 se observa la arquitectura estándar de una GRU incluye dos compuertas principales. La compuerta de actualización (update gate) Controla cuánto de la información anterior se mantiene en el estado actual, permitiendo al modelo retener información relevante a lo largo del tiempo. La compuerta de reinicio (reset gate): Determina cuánta información previa debe olvidarse, facilitando la incorporación de nueva información relevante [49].

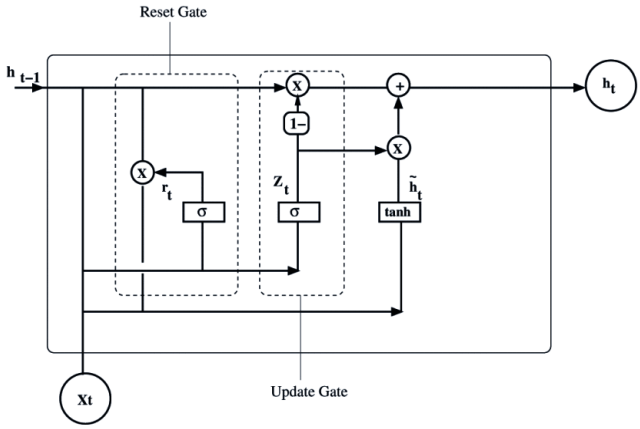

Figura 2.5 Estructura básica del modelo GRU.

Fuente: [11].

2.2.3.9 Optimización con Optuna

La optimización de hiperparámetros en redes recurrentes resulta esencial para garantizar un buen desempeño en tareas de clasificación secuencial. Parámetros como el número de neuronas, la tasa de aprendizaje o los mecanismos de regularización determinan tanto la capacidad de representación del modelo como su capacidad de generalización.

- Número de neuronas en cada capa LSTM: selecciona la cantidad de neuronas en el rango de 32 a 256 [17].
- Learning rate: Selecciona una taza de aprendizaje entre 1e-4 a 1e-2[11].
- Número de capas: Busca el número de capas LSTM necesarias entre 2 y 4 para capturar las secuencias [17].
- Dropout: regularización para prevenir el sobreajuste entre 0.1 y 0.6 con pasos de 0.1.
- Regularización: Regularización L2 para evitar el sobreajuste, garantizando una buena generalización.
- Unidades de capas densas: Se optimiza el número de unidades en las capas densas completamente conectadas, con rango entre 32 a 128 [17].

2.2.4 Vibraciones acústicas

Las vibraciones acústicas son oscilaciones mecánicas de partículas u otro medio elástico que se propagan como ondas sonoras. Estas vibraciones son percibidas por el cuerpo humano como sonido. Cuando se emite energía acústica, causa una perturbación cercana a la fuente. Esta perturbación se transmite como una serie de compresiones y rarefacciones sucesivas en el medio, formando una onda de presión, a medida que estas ondas de presión se propagan, las partículas del medio oscilan alrededor de su posición de equilibrio, generando vibraciones acústicas donde son la base de las ondas sonoras, las cuales se generan cuando las partículas de un medio como el aire se desplazan hacia adelante y hacia atrás debido a una perturbación, creando una propagación de compresiones y rarefacciones. Algunos aspectos claves de las vibraciones acústicas son la frecuencia, amplitud, longitud de onda y resonancia y tienen diversas aplicaciones, como la reproducción de música, la comunicación verbal, tecnologías de ultrasonido en medicina, sonar en navegación, control de ruido en ingeniería acústica y más. [50].

2.2.4.1 Filtros Gammatone (GTF)

Los filtros Gammatone son modelos matemáticos diseñados para imitar como la membrana basilar de la cóclea humana responde a diferentes frecuencias del sonido, específicamente de las células ciliadas en la membrana basilar del oído interno. En la

Figura 2.6 se puede observar la estructura del oído interno, destacando los componentes clave involucrados en el procesamiento auditivo.

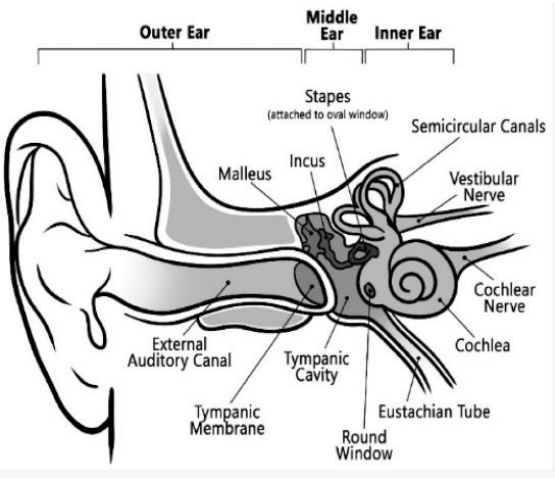

Figura 2.6 Componentes del oído humano, El sonido entra en el oído externo a través del pabellón auricular y viaja hasta el oído medio e interno. Finalmente, llega a la cóclea y hace vibrar la membrana basilar.

Fuente: [51].

Son filtros pasa banda cuya función de respuesta al impulso es una modulación de una función Gamma con una onda portadora senoidal:

$$h(t) = ct^{n-1}e^{-2\pi bt}\,cos(2\pi f_c t + \emptyset)\,u(t) \qquad \textit{Ecuación 2.6}$$

Donde c es la constante de proporcionalidad, n el orden del filtro, b coeficiente de decaimiento temporal, f_c frecuencia portadora, $\emptyset$ fase de la portadora y $u(t)$ función escalón.

La composición del filtro se divide en dos partes, parte temporal $t^{n-1}e^{-2\pi bt}$ modela como decae la respuesta del oído a lo largo del tiempo, parte cosenoidal $\cos(2\pi f_c t + \emptyset)$ es la encargada de introducir la selectividad de frecuencia.

Los GTF permiten descomponer la señal centrándose en frecuencias criticas distintas, y al combinarse en un banco de filtros, permite dividir una señal de audio en componentes de frecuencia similar a los del oído humano [51].

2.2.4.2 Transformada Hilbert

La transformada Hilbert permite extraer la información de fase y envolvente de señales reales, siendo una herramienta clave en el análisis de señales donde no cambia el contenido espectral de la señal, introduce un desfase de -90° en cada componente de frecuencia [52]. La transformada Hilbert se define en la ecuación 2.7:

$$\hat{x}(n) = x(n) * h(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) \qquad \textit{Ecuación 2.7}$$

Donde $h(n) = \begin{cases} \frac{2}{\pi n}, & n\ impar \\ 0, & n\ par \end{cases}$

Junto con la señal original, la transformada Hilbert permite construir la señal analítica

$$z(n) = x(n) + j\hat{x}(n) \qquad \textit{Ecuación 2.8}$$

Con esta señal se puede obtener la envolvente, calcular la fase instantánea y extraer la frecuencia instantánea (ecuaciones 2.9, 2.10, 2.11).

$$Envolvente: A(n) = |z(n)| = \sqrt{x(n)^2 + \hat{x}(n)^2} \qquad \textit{Ecuación 2.9}$$

$$Fase\ instantane: \phi(n) = arg\big(z(n)\big) = arctan\ (\frac{\hat{x}(n)}{x(n)}) \qquad \textit{Ecuación 2.10}$$

$$Frecuencia\ instantanea: f_{Hz}(n) = \frac{f(n)fs}{2\pi} \qquad \textit{Ecuación 2.11}$$

2.2.5 Indicadores gráficos

2.2.5.1 Formantes

Los formantes son frecuencias resonantes del tracto vocal que caracterizan los sonidos del habla. Su estimación precisa es esencial en áreas como el reconocimiento de voz, la síntesis de habla y el análisis fonético. Se dividen en dos formantes principales: Primer formante (F1) está relacionado con la apertura de la boca. Una boca más abierta tiene mayor frecuencia de F1, una boca más cerrada tiene menor frecuencia de F1. Este formante típicamente se encuentra en el rango de 300 Hz a 1000 Hz [53].

Segundo formante (F2) está relacionado con la posición de la lengua, si la lengua esta más hacia adelante tiene mayor frecuencia de F2. Si la lengua esta más hacia atrás tiene menor frecuencia de F2 [54], se encuentra en el rango de 1000 Hz a 3000 Hz [53].

2.2.5.2 Pitch

El pitch es una característica clave en el análisis de señales de voz, relacionada con la frecuencia fundamental de la oscilación glotal. Es fundamental para síntesis de voz, identificación de variaciones en la entonación, acentuación y emociones del hablante. El pitch varía a lo largo del tiempo debido a la naturaleza cuasi-periódica de las oscilaciones glotales y la articulación del tracto vocal [55].

2.2.6 Métricas de desempeño

Las métricas de desempeño son medidas cuantitativas que se utilizan para evaluar la efectividad y eficiencia de un sistema, proceso o modelo en función de ciertos criterios u objetivos predefinidos. Estas métricas permiten cuantificar y analizar diversos aspectos del rendimiento, proporcionando una base objetiva para la toma de decisiones

y la mejora continua. Las métricas de desempeño varían según el contexto y el dominio de aplicación [56].

2.2.6.1 Matriz de confusión

La matriz de confusión es una herramienta que se utiliza para evaluar el rendimiento de un clasificador que ha sido entrenado con técnicas de aprendizaje supervisado. Esta herramienta proporciona información cuantitativa sobre el número de aciertos y errores de clasificación para cada una de las clases. A partir de la información que se obtiene de la matriz de confusión, se pueden calcular diversas métricas de desempeño [57].

- TP (True Positives): número de predicciones correctas de la clase Positive. También se denomina Potencia (Power).
- FN (False Negatives): número de falsas predicciones de la clase Negative, cuando realmente son de la clase Positive. También se denomina errores tipo II.
- FP (False Positives): número de falsas predicciones de la clase Positive, cuando realmente son de la clase Negative. También se denomina errores tipo I.
- TN (True Negatives): número de predicciones correctas de la clase Negative

2.2.6.2 Exactitud

Se refiere al grado en que el resultado de una medición/predicción se ajusta al valor correcto. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación y forma práctica la exactitud es la cantidad total de predicciones que fueron correctas. Es una métrica general, pero puede no ser suficiente en entornos desbalanceados [56].

$$ACC = \frac{TP+TN}{P+N} \qquad \textit{Ecuación 2.12}$$

2.2.6.3 Precisión

La precisión o Confianza denota la proporción de casos positivos predichos que son correctamente positivos Reales. Se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Está relacionado al porcentaje de casos positivos detectados y es útil en data sets balanceados [56].

$$PPV = \frac{TP}{TP+FP} \qquad \textit{Ecuación 2.13}$$

2.2.6.4 Recall o sensibilidad

Indica la proporción de instancias positivas correctamente identificadas entre todas las

instancias positivas reales. También se conoce como tasa de Verdaderos Positivos (True Positive Rate). Es la fracción de casos positivos que fueron correctamente identificadas por el algoritmo [57].

$$TPR = \frac{TP}{RP} \qquad \textit{Ecuación 2.14}$$

2.2.6.5 F1 Score

Es la media armónica de precisión y sensibilidad, buscando un equilibrio entre ambas métricas, no considera los verdaderos negativos y su utilidad prevalece cuando la distribución de clases es desigual y se enfrenta a desbalance de datos [57].

$$F1 = 2\left(\frac{PPV * TPR}{PPV + TPR}\right) \qquad \textit{Ecuación 2.15}$$

2.2.7 Resumen

En este capítulo se presentaron los fundamentos teóricos con base a la literatura, necesarios para la construcción de un sistema audiovisual de ayuda para personas sordas. Se abordaron conceptos claves como es el procesamiento de señales, las características de una señal de voz, los diferentes tipos de clasificadores de aprendizaje automático, los indicadores gráficos de una comparación de dos señales de audio y los conceptos para la generación de vibraciones acústicas.

En el procesamiento de señales se abarca el concepto de transformada de Fourier para análisis en dominio frecuencial, la eliminación de ruido mediante filtros pasa altas y sustracción espectral además del uso de la ventana Hanning para segmentar una señal de audio en diferentes frames temporales. Así mismo, se destacaron características como los coeficientes cepstrales de frecuencia de Mel (MFCCs), vectores chroma y el espectrograma, además se hizo énfasis en el uso de modelos avanzados como wav2vec 2.0, que permite la extracción de características directamente de datos crudos.

Para la etapa de reconocimiento de voz, se consideraron clasificadores basados en modelos ocultos de Markov (HMM) y redes neuronales profundas, como LSTM y GRU. Para la generación de señales vibro táctiles se analiza el uso de los filtros Gammatone (GTF) y la transformada Hilbert, los cuales constituyen componentes metodológicos fundamentales para el desarrollo del sistema propuesto.

3. Metodología para el desarrollo del sistema audiovisual de ayuda para personas sordas

Se describe un sistema audiovisual y vibrotáctil de ayuda para la rehabilitación del habla, específicamente para asistir a personas con dificultades en la producción del lenguaje oral. La Figura 3.1 presenta un diagrama de bloques que ilustra la solución propuesta. En la primera etapa, se realiza la captura de la señal a través de un micrófono. La segunda etapa es el procesamiento, que incluye técnicas de filtrado. La tercera etapa es de clasificación consta de dos métodos principales:

1) Método de extracción de características predefinidas: La señal del audio es preprocesada y se extraen características. Estas características se entregan a los modelos de clasificación LSTM y HMM.

2) Procesamiento directo con Transformers: La señal de audio se pasa directamente en crudo al modelo pre entrenado wav2vec 2.0, que se encarga de la extracción automática de características. Los resultados de la esta extracción son procesados por los clasificadores LSTM y GRU

Una vez completado el proceso de clasificación a partir del modelo basado en Wav2vec 2.0, el resultado final es presentado al usuario a través de una interfaz interactiva. En esta interfaz, el usuario puede observar la palabra obtenido en la clasificación, visualizar un indicador de pronunciación y recibir retroalimentación mediante vibraciones acústicas, mejorando así la experiencia.

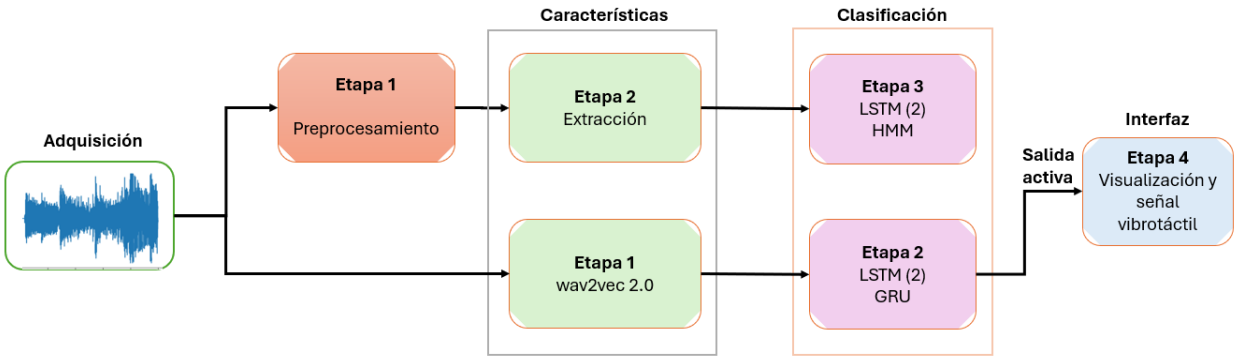

Figura 3.1 Diagrama de bloques de la solución propuesta.

Fuente: Elaboración propia.

3.1 Descripción del dataset

El conjunto de datos *OpenSLR SLR72* [58] constituye un recurso valioso para el desarrollo y evaluación de sistemas de procesamiento del lenguaje natural orientados al español colombiano. Este corpus, compilado y liberado por Google Inc. bajo la licencia Creative Commons Atribución-CompartirIgual 4.0 Internacional, contiene grabaciones de voz de alta calidad producidas por hablantes voluntarios, organizadas

según el género de los participantes y representa un insumo esencial para avanzar en la personalización y adaptación de tecnologías del habla al contexto colombiano, permitiendo el entrenamiento y validación de modelos que reconozcan las particularidades fonéticas y prosódicas del español hablado en Colombia. El dataset cuenta con 2537 grabaciones pronunciadas por 16 hombres y 2372 grabaciones pronunciadas por 15 mujeres, ambos con acento colombiano de Bogotá, la frecuencia de grabación fue de 48 KHz lo cual permite obtener una mayor observación de las características de las grabaciones.

La selección de palabras para el sistema se realizó con base en criterios fonéticos definidos en la sección 2.2.1.2, los cuales agrupan los sonidos del habla según su punto de articulación: alveolares, bilabiales, palatales y velares. Se identificaron las palabras que contienen fonemas representativos de cada grupo, permitiendo una cobertura articulatoria equilibrada. Además del criterio fonético, se consideró la disponibilidad de muestras en el dataset SLR72, solo se incluyeron palabras que, además de cumplir con la clasificación articulatoria, presentaban una cantidad suficiente de ejemplos para garantizar un entrenamiento robusto y una evaluación representativa del modelo. Como resultado, se seleccionaron tres palabras por grupo fonético, priorizando aquellas con mayor presencia en el corpus disponible, tal como se resume en la Tabla 3.1.

Tabla 3.1Palabras seleccionadas del dataset SLR72.

Fuente: Elaboración propia.

Alveolares	**N° Muestras**	**Bilabiales**	**N° Muestras**	**Palatales**	**N° Muestras**	**Velares**	**N° Muestras**
Casa	47 F 56 M	Mucho	54 F 54 F	Lluvia	19 F 22 M	Agua	14 F 19 M
Para	41 F 41 M	Peso	19 F 19 M	Calle	36 F 38 M	Amigo	24 F 25 M
Sol	45 F 30 M	Bicicleta	34 F 34 M	Noche	21 F 20 M	Tarjeta	13 F 14 M

Cada una de las grabaciones seleccionadas tienen una duración entre 500 milisegundos y 1 segundo, además no presentan ruido o interferencias y no es necesario aplicar filtrado en cada señal.

Aumento de datos: El aumento es esencial en el procesamiento de señales, especialmente cuando se trabaja con bases de datos desbalanceados, en la Tabla 3.1 se puede observar la diferencia de muestras para cada clase, en este caso las clases minoritarias están subrepresentadas, lo que puede sesgar el modelo hacia clases mayoritarias y reducir su capacidad de generalización. Para este conjunto de datos se aplicaron técnicas como: la adición de ruido blanco que consiste en añadir una señal aleatoria con igual intensidad en todas las frecuencias a la señal original como se evidencia en la Figura 3.2, cambio de pitch para aumentar el tono sin cambiar la duración, cambio de tempo para modificar la velocidad de la señal, pero sin alterar su tono y el desplazamiento temporal hacia adelante o atrás [18], agregando un desfase que no afecte a las características principales de la señal.

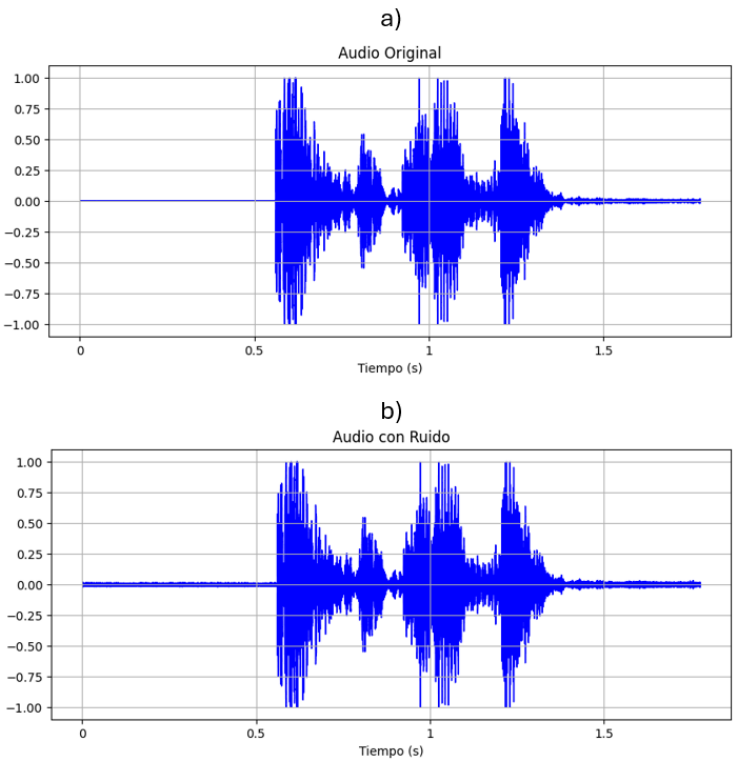

Figura 3.2 Señal de audio de la palabra bicicleta. a) señal original; b) señal con ruido blanco agregado.

Fuente: Elaboración propia.

El criterio para seleccionar la cantidad y tipo de técnica de aumento de datos aplicada a cada clase con base en obtener un dataset balanceado. La clase con mayor cantidad de audios fue Mucho, con 108, y la clase con menor cantidad fue Tarjeta, con 27 audios. A partir de esta distribución, se divide en 70% para entrenamiento y 30% para test. En la Figura 3.3 puede apreciarse el aumento de datos solo en los datos de entrenamiento aumentando las grabaciones a la clase que cuenta con más datos.

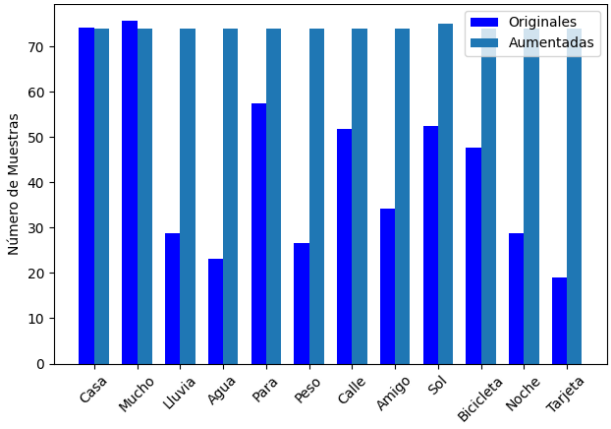

Figura 3.3 Distribución de audios por clases después del aumento de datos.

Fuente: Elaboración propia.

3.2 Adquisición voz del usuario

En la etapa de adquisición, el sistema captura la señal de voz del usuario a través de la interfaz utilizando un micrófono como dispositivo, se usa una frecuencia de muestreo de 16000 Hz que permite capturar adecuadamente el rango útil del habla (0–8 kHz), el canal mono es suficiente al no requerirse información espacial del sonido y los 16 bits garantizan un rango dinámico adecuado para representar la voz con buena calidad [15]. Parámetros que son estándar en sistemas de reconocimiento de voz. Esta señal de voz es la entrada principal sobre la cual operan las demás etapas. La Figura 3.4 muestra la señal de voz del usuario adquirida.

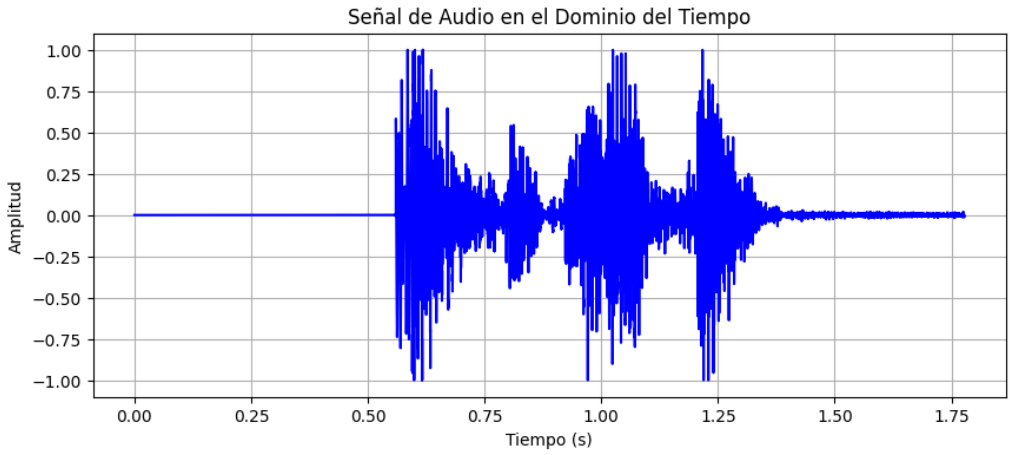

Figura 3.4 Señal de audio de la palabra bicicleta.

Fuente: Elaboración propia.

3.3 Procesamiento de la señal

Esta etapa tiene como finalidad adecuar la señal de voz mediante métodos de filtrado y normalización. Una vez adquirida la señal de voz debe pasar por la etapa de procesamiento, durante el procesamiento, se aplican técnicas para filtrar el ruido por medio de un filtro pasa altas, sustracción espectral, posteriormente se normaliza. En la Figura 3.5 se ilustra el diagrama de bloques para la etapa de procesamiento de la señal.

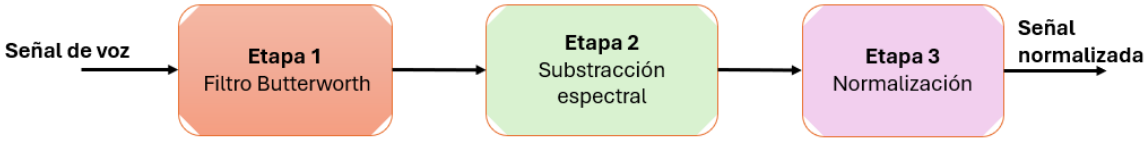

Figura 3.5 Diagrama de bloques para la etapa de procesamiento de la señal.

Fuente: Elaboración propia.

Filtrado: El procedimiento de filtrado diseñado consiste en aplicar un filtro digital pasa-altas Butterworth de quinto orden [33] con una frecuencia de corte de 120 Hz [32], adecuado para eliminar componentes de baja frecuencia (<100-120Hz) que no contienen información relevante del habla, como zumbidos eléctricos (frecuencia de 60 Hz de la red eléctrica), vibraciones o ruido de fondo grave sin afectar la discriminación lingüística. Para asegurar una salida sin desfase de fase el filtro se aplica en ambas direcciones. Con este tipo de filtro Butterworth se obtiene una supresión efectiva del ruido de baja frecuencia, sin comprometer la estabilidad numérica ni incrementar el costo computacional, asegurando una respuesta suave y sin ondulaciones en la banda pasante permitiendo una transición natural sin distorsiones abruptas. El espectrograma de la señal en la Figura 3.6b muestra el resultado de aplicar el filtro.

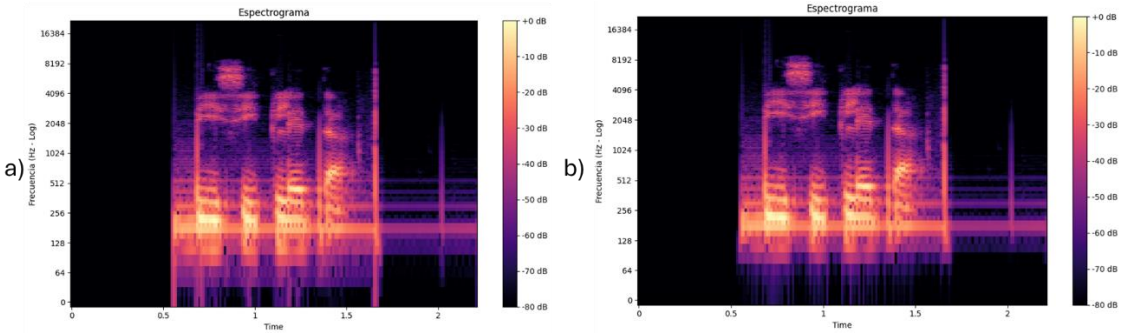

Figura 3.6 Espectro de la señal de la palabra bicicleta. a) señal antes de aplicar el filtro Butterworth; b) Señal después de aplicar el filtro Butterworth.

Fuente: Elaboración propia.

Sustracción espectral: Para la realización del filtrado por sustracción espectral se requiere tomar una muestra del ruido presente en la señal. Esta muestra debe ser capturada durante un segmento de silencio al inicio de la grabación sin tomar parte de la señal útil, creando un perfil de ruido y aplicar la transformada de Fourier en corto tiempo (STFT) Figura 3.7 para dividir la señal en ventanas para calcular la frecuencia presente en cada ventana y obtener el espectrograma.

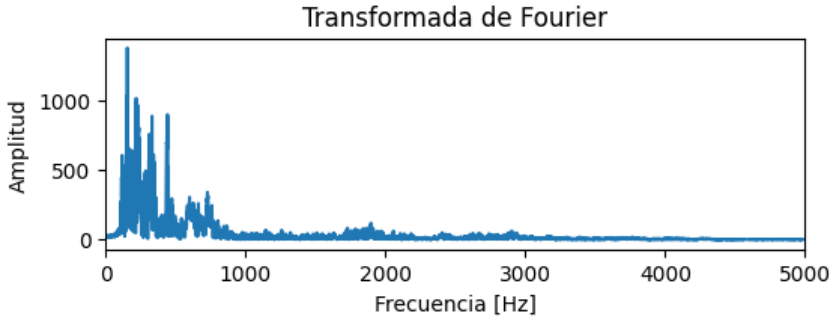

Figura 3.7 Transformada de Fourier de la señal de audio antes de usar el filtrado por sustracción espectral.

Fuente: Elaboración propia.

El perfil de ruido se resta a las magnitudes del espectrograma de la señal completa, si el nivel de energía en una frecuencia es similar al nivel de ruido, esa frecuencia se atenúa. Si el nivel de energía es significativamente mayor que el del ruido, se considera que contiene información útil y se conserva la señal, después de eliminar el ruido se aplica la transformada de Fourier inversa para reconstruir la señal. En la Figura 3.8 se puede observar la comparación de los espectrogramas donde muestra una mejora significativa en la señal tras aplicar la sustracción espectral. En el espectrograma de la Figura 3.8a, se observa un ruido de fondo, predominante en las frecuencias bajas entre 0 a 500 Hz, lo que afecta la claridad de frecuencias útiles entre 500 Hz y 3 kHz. En el espectrograma de la Figura 3.8b se observa la señal procesada, el ruido de fondo ha sido atenuado, mejorando la relación señal-ruido. Las frecuencias relevantes permanecen claras, mientras que los componentes por encima de los 3 kHz también fueron atenuados, eliminando el ruido no deseado sin afectar la calidad de la señal, preservando las características esenciales de la señal de voz.

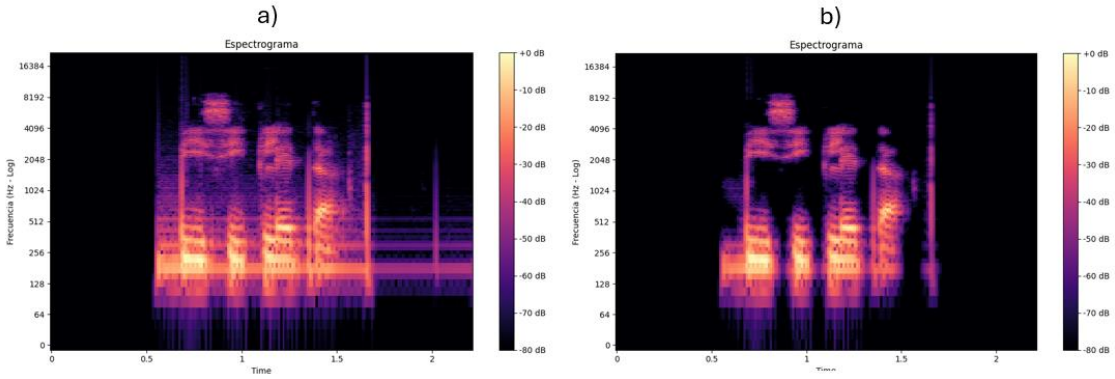

Figura 3.8 Espectro de la señal para la palabra bicicleta. a) señal antes de aplicar la reducción espectral; b) señal después de aplicar la reducción espectral.

Fuente: Elaboración propia.

En la Figura 3.9 se observa la transformada de Fourier de la señal ya procesada, se evidencia la efectividad del procesamiento para la reducción del ruido en frecuencias bajas y altas a comparación de la Figura 3.7.

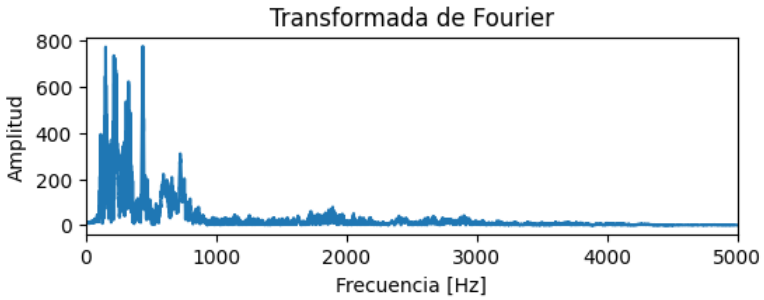

Figura 3.9 Transformada de Fourier de la señal de audio post filtrado.

Fuente: Elaboración propia.

La señal de salida del filtro es normalizada de [-1, 1]. Para ello, se calcula valores mínimos y máximos absolutos de amplitud, sin distinguir entre signos positivos o negativos. Los cuales se usan para restar el valor mínimo a la señal y dividir entre la diferencia entre máximo y el mínimo, este procedimiento permite ajustar las amplitudes de la señal, evitando distorsiones y garantizando una representación adecuada sin alterar su contenido frecuencial ni su duración temporal. Se optó por el rango debido a que es un estándar ampliamente utilizado en el procesamiento digital de señales de audio, ya que facilita la representación numérica en sistemas de punto flotante, evita y mejora la estabilidad numérica de algoritmos de procesamiento [30], al mantener las amplitudes en un intervalo acotado y simétrico alrededor de cero. En la Figura 3.10 se presenta la señal de audio resultante después del proceso de normalización.

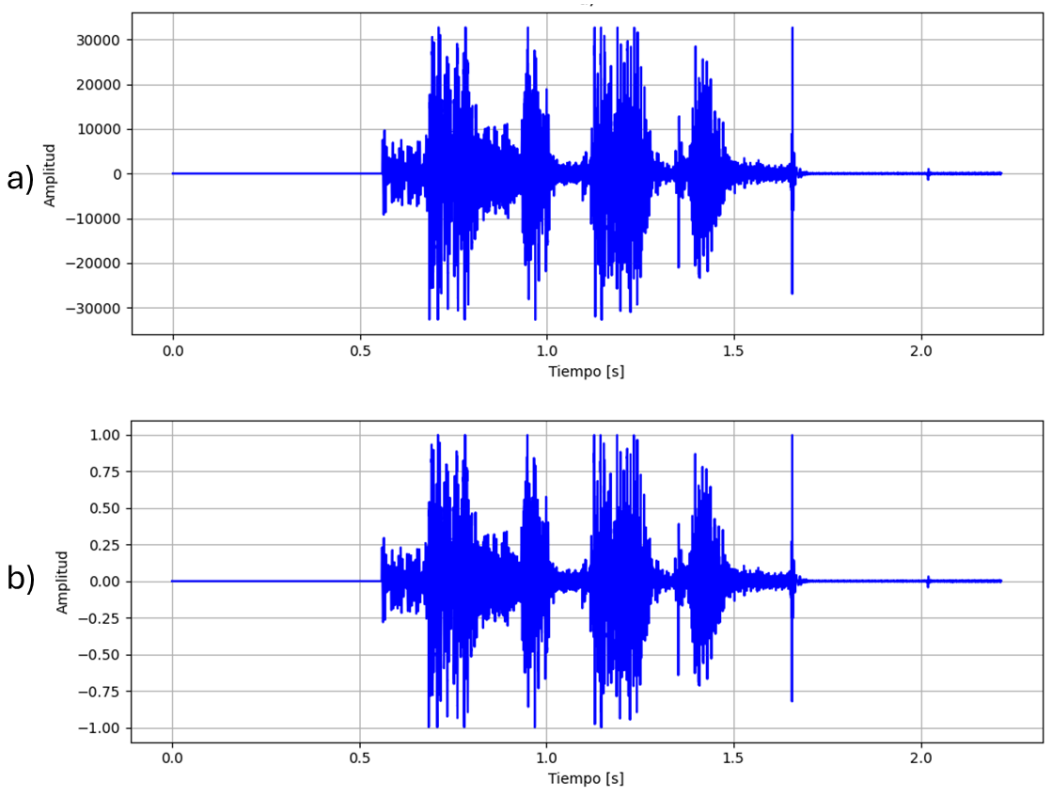

Figura 3.10 Señal de audio de la palabra bicicleta. a) señal sin normalizar; b) señal normalizada.

Fuente: Elaboración propia.

3.4 Método de extracción de características predefinidas

3.4.1 Segmentación

El proceso de segmentación se implementó para analizar las grabaciones en ventanas temporales discretas y conservar las características temporales de las señales. Cada grabación fue dividida en segmentos de 50 ms con un solapamiento de 50% entre cada segmento [9], [16] como se presenta en la Figura 3.11a, esto para asegurar una continuidad en la información. A cada segmento se le aplico una ventana Hanning,

ideal para reducir las discontinuidades en los bordes de los segmentos al suavizar gradualmente el inicio y fin de cada ventana [35] Figura 3.11b. Esta técnica garantiza que las características frecuenciales extraídas de cada segmento reflejen fielmente las propiedades acústicas del audio.

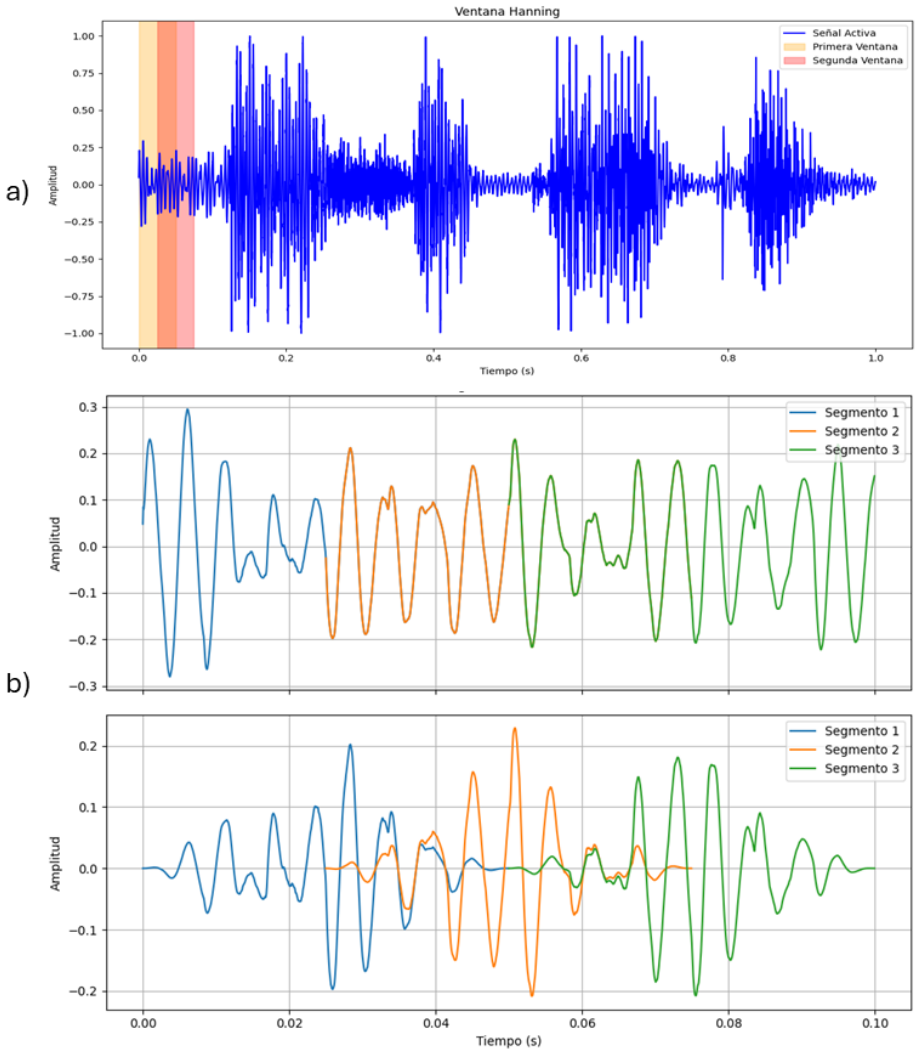

Figura 3.11 Señal de voz de la palabra bicicleta. a) ejemplo de segmentación de la señal para dos ventanas; b) primeros tres segmentos antes y después de aplicar la ventana Hanning.

Fuente: Elaboración propia.

3.4.2 Extracción de características

Esta etapa tiene como finalidad la extracción de características de las señales de audio para obtener un vector de características que describa a detalle las señales de voz del conjunto de datos. Se aplicó un enventanado a cada muestra para mejorar la representación temporal.

MFCC (Mel Frequency Cepstral Coefficients) y derivadas: Para cada grabación de audio, se extrajeron los coeficientes cepstrales en las frecuencias de Mel (MFCC), junto con sus primeras y segundas derivadas temporales (delta y delta-delta) [9]. El procedimiento se aplicó sobre segmentos solapados de la señal utilizando una transformada de Fourier. Se utilizaron 40 filtros Mel para calcular las energías espectrales, y se estableció un límite superior de frecuencia de 7000 Hz [16] para captar la banda útil del habla.

A partir de estas configuraciones, se obtuvieron 13 coeficientes MFCC por cada segmento, que luego se complementaron con sus derivadas de primer y segundo orden, formando un vector de características de 39 dimensiones por ventana temporal. Estas secuencias de vectores se mantuvieron en su forma original sin promediar, preservando así la dinámica temporal del habla. Cada muestra fue finalmente representada por una matriz de tamaño variable (T, 39), donde T corresponde al número de ventanas de la señal de audio.

Espectrograma: El espectrograma se calculó aplicando la Transformada de Fourier de corto plazo (STFT) para cada segmento y se obtuvo el espectro de magnitud [13], posteriormente se elevó al cuadrado para representar la distribución de energía en el dominio tiempo-frecuencia, generando una matriz por cada muestra.

Chroma: Las características Chroma se extrajeron a partir del espectro de magnitud obtenido con la STFT. Las frecuencias fueron mapeadas a 12 bins correspondientes a los semitonos de la escala cromática [47], acumulando la energía de todas las frecuencias que pertenecen a cada clase, obteniendo una representación (T, 12) por muestra, donde T es el número de ventanas temporales.

Normalización: Después de extraer las características se aplicó una normalización mediante Z-Score, estandarizando cada dimensión del vector de características para que tuviera media cero y desviación estándar uno. Este procedimiento se realizó de forma independiente para cada conjunto de datos (entrenamiento y prueba) utilizando los parámetros calculados sobre el conjunto de entrenamiento, con el fin de evitar filtraciones de información. La normalización Z-Score fue seleccionada por ser una técnica ampliamente utilizada en el procesamiento de señales de voz, ya que mejora la estabilidad numérica, acelera la convergencia de los algoritmos de aprendizaje y evita que las características con mayores magnitudes dominen el entrenamiento del modelo [36].

En la Figura 3.12 se observa los MFCC de una grabación, la Figura 3.12a muestra los coeficientes sin normalizar, denotando una amplitud significativamente diferente, algunos coeficientes presentan valores muy grandes. La Figura 3.12b después de normalizar con Z-Score todos los coeficientes tienen una distribución centrada alrededor de 0 y una desviación estándar de aproximadamente 1, donde los outliers son más visibles en todos los coeficientes.

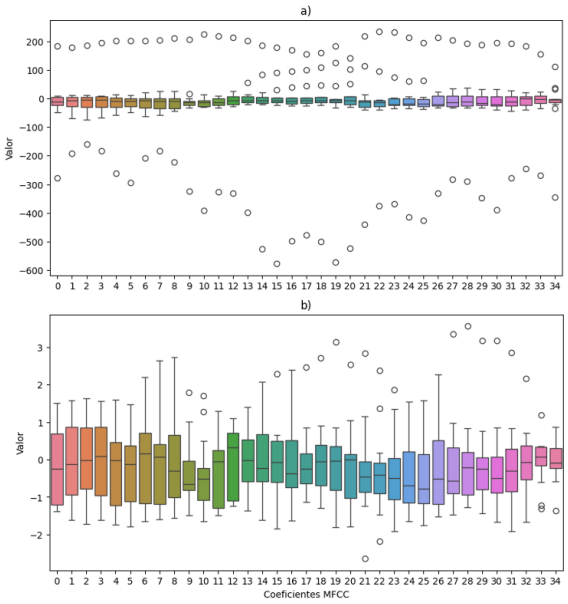

Figura 3.12 Coeficientes MFCC. a) Coeficientes MFCC sin normalizar. b) Coeficientes MFCC con normalización Z-Score.

Fuente: Elaboración propia.

3.4.3 Clasificación

Inicialmente, se realizó una división estratificada del conjunto de datos Tabla 3.1, asignando el 70% de los datos para entrenamiento y el 30% para prueba, asegurando que las grabaciones de una misma palabra pronunciada por un usuario no estuvieran simultáneamente en los conjuntos de entrenamiento y prueba. Debido al desbalance en el conjunto de datos, se aplicó un proceso de aumento de datos sección 3.1 para mejorar la representatividad de las clases minoritarias, logrando una distribución suficiente de datos para el aprendizaje del modelo, manteniendo al mismo tiempo un conjunto de prueba representativo y estadísticamente robusto para la evaluación final.

Para la selección de hiperparámetros, se priorizó la métrica de precisión, ya que en este problema resulta fundamental minimizar los falsos positivos, asegurando que las predicciones positivas correspondan efectivamente a la clase correcta, lo cual es especialmente relevante en escenarios desbalanceados donde métricas como la exactitud (accuracy) pueden resultar engañosas. A continuación, se presenta una descripción de cada método, su configuración y los procedimientos empleados.

3.4.3.1 Modelos ocultos de Márkov (HMM)

Se implementaron Modelos Ocultos de Márkov (HMM) con observaciones continuas multivariadas, estas observaciones se agruparon en secuencias por palabra para el entrenamiento de modelos independientes por clase.

Estados visibles: En este caso, se utilizó la concatenación de 13 coeficientes MFCC con sus respectivas primeras y segundas derivadas temporales (delta y delta-delta),

espectrograma y características chroma, resultando en un vector de 39 dimensiones por frame. Estas observaciones se agrupan en secuencias por palabra.

Número de estados (n_components): Para optimizar la capacidad representacional de los modelos, se asignó un número distinto de estados ocultos a ciertas palabras según su complejidad fonética. Se utilizaron 4 estados para palabras de dos sílabas, 5 estados para “Tarjeta” (tres sílabas) y 6 estados para “Bicicleta” (cuatro sílabas), permitiendo al modelo capturar secuencias acústicas más complejas.

Matriz de transición (transmat_): Fue inicializada de forma uniforme para evitar sesgos de entrada, y posteriormente suavizada mediante una regularización L2 con un factor lambda_reg = 0.01. Luego se normalizó fila por fila para garantizar que las probabilidades de transición entre estados sumaran uno, reflejando trayectorias plausibles de producción del habla.

Distribuciones de emisión (covariance_type= 'diag'**)**: Se asumieron covarianzas diagonales, lo que reduce la complejidad computacional y evitar el sobreajuste.

Algoritmo de entrenamiento: Se utilizó el algoritmo Baum-Welch, una forma del Expectation-Maximization (EM), que estima de manera iterativa los parámetros del modelo a partir de las secuencias de entrenamiento. Este algoritmo fue elegido por ser el método estándar para entrenar HMM cuando los estados ocultos no son observables directamente. El entrenamiento se llevó a cabo por un máximo de n_iter = 100 iteraciones con parada temprana por convergencia, este valor sigue los límites por defecto usados en implementaciones de referencia del algoritmo y evita abortar prematuramente entrenamientos que requieren múltiples re-estimaciones para estabilizarse.

Inicialización (init_params = 'smc'**):** Esta opción permite inicializar los parámetros del modelo (startprob, means, covars) de forma aleatoria, excepto la matriz de transición, que fue establecida explícitamente.

3.4.3.2 LSTM (long short term memory)

El conjunto de datos se dividió en 70% para entrenamiento y 30% para prueba. Se definieron dos configuraciones de LSTM (**LSTM1 y LSTM2**) y cada arquitectura del modelo fue optimizada mediante Optuna de Python, encargado de buscar y seleccionar los hiperparámetros para maximizar el rendimiento del modelo, priorizando la métrica precisión. Los rangos de búsqueda fueron establecidos considerando los valores comúnmente reportados en la literatura donde especifica que el rendimiento de las redes LSTM depende de una selección adecuada de sus parámetros estructurales y de entrenamiento.

Optimización con Optuna modelo LSTM1

La exploración determinó que la configuración óptima consiste en tres capas LSTM apiladas de 160 neuronas cada una, seguidas por una capa densa con activación

Softmax y una tasa de aprendizaje de 0.00164, valor que proporcionó el mejor equilibrio entre precisión y estabilidad durante el entrenamiento.

La primera capa de 160 neuronas recibe como entrada una secuencia de longitud fija T (número de frames) con F características por paso temporal. Esta capa es bidireccional que permite capturar tanto pasado como futuro dentro de la secuencia, lo que mejora la comprensión de la estructura temporal global. Su objetivo es capturar patrones locales de variación temporal en los coeficientes MFCC y sus derivadas (deltas). Al devolver una secuencia de salidas del mismo largo que la entrada, esta capa mantiene la estructura temporal para su procesamiento posterior.

La segunda capa de 160 neuronas bidireccional toma como entrada la secuencia generada por la capa anterior y profundiza el análisis temporal, permitiendo al modelo aprender dependencias de mayor alcance entre los elementos de la secuencia. Esta arquitectura jerárquica ayuda a identificar patrones fonéticos más complejos y distribuidos a lo largo del tiempo.

La tercera capa resume la información secuencial aprendida en las capas anteriores, generando un vector de estado oculto que representa la secuencia completa. Esta vectorización final es esencial para alimentar la capa densa de clasificación, ya que proporciona una representación compacta y discriminativa de la señal acústica.

Por último, la capa densa con función de activación softmax [17] y recibe el vector resumen de la tercera capa LSTM donde produce una distribución de probabilidad sobre las posibles clases (palabras). El modelo se entrena con la función de pérdida entropía cruzada categórica dispersa (sparse_categorical_crossentropy), adecuada para tareas de clasificación multiclase con etiquetas enteras.

Además, el modelo aplica dropout de 0.1232 después de cada capa LSTM. Este método de regularización desconecta aleatoriamente un porcentaje de unidades durante el entrenamiento, lo que obliga al modelo a no depender excesivamente de ninguna neurona en particular. Se incorpora regularización L2 de 4.81e-6 en cada capa LSTM y en la capa densa. Esta penalización sobre los pesos grandes del modelo ayuda a mantener la complejidad del modelo bajo control, fomentando soluciones más robustas y generalizables.

Optimización con Optuna modelo LSTM2

La exploración determino una estructura general del modelo se compone de 2 capas LSTM apiladas, seguidas por una capa densa de clasificación y una taza de aprendizaje de 0.00179. El modelo inicia con una capa de 96 neuronas bidireccional con una entrada de tamaño fija para aprender las secuencias en ambos sentidos y entregar una salida del mismo tamaño de la entrada.

La segunda capa de 96 neuronas bidireccional toma como entrada la secuencia generada por la capa anterior y profundiza el análisis temporal, permitiendo al modelo aprender dependencias de mayor alcance entre los elementos de la secuencia

Por último, la capa densa con función de activación softmax [17] y recibe el vector resumen de la segunda capa LSTM donde produce una distribución de probabilidad sobre las posibles clases (palabras). Además, se aplica dropout de 0.3392 después de cada capa LSTM. Este método de regularización desconecta aleatoriamente un porcentaje de unidades durante el entrenamiento, lo que obliga al modelo a no depender excesivamente de ninguna neurona en particular. Se incorpora regularización L2 de 3.299e-05en cada capa LSTM y en la capa densa.

3.5 Extracción de características con Transformers

3.5.1 Extracción de características con wav2vec 2.0

Se realizó una división estratificada del conjunto de datos Tabla 3.1, asignando el 70% de los datos para entrenamiento y el 30% para prueba, asegurando que las grabaciones de una misma palabra pronunciada por un usuario no estuvieran simultáneamente en los conjuntos de entrenamiento y prueba. Debido al desbalance en el conjunto de datos, se aplicó un proceso de aumento de datos Sección 3.4 para posteriormente pasar las grabaciones al modelo wav2vec 2.0.

El modelo preentrenado wav2vec 2.0 permite la extracción de características de señales de voz directamente desde datos crudos. Para procesar todas las grabaciones del conjunto de datos, se ajustó la frecuencia de muestreo a 16 kHz, requisito necesario para su compatibilidad con el modelo. Posteriormente, las señales fueron procesadas a través del encoder, donde se aplicaron capas convolucionales destinadas a capturar patrones de baja frecuencia, generando un conjunto de representaciones comprimidas. Estas representaciones se ingresaron al contextualizador, que utiliza un Transformer para modelar las relaciones entre las representaciones comprimidas. El resultado de este proceso son embeddings que encapsulan las características acústicas, a las cuales se les aplicó reducción de dimensionalidad mediante PCA, optimizando así el uso de recursos computacionales.

3.5.2 Clasificación

Para el entrenamiento de los modelos, la métrica priorizada para seleccionar los hiperparámetros fue precisión, con el fin de medir la proporción de verdaderos positivos entre todos los casos predichos como positivos. A continuación, se presenta una descripción de cada método, su configuración y los procedimientos empleados para la búsqueda de hiperparámetros con Optuna de Python.

3.5.2.1 LSTM (Long Short Term-Memory) y GRU (Gated Recurrent Unit)

Una vez divido el conjunto de datos, se definieron dos configuraciones de LSTM (LSTM_T1 y LSTM_T2) y GRU, para cada configuración se usó Optuna para encontrar la mejor combinación de hiperparámetros.

Optimización con Optuna modelo LSTM_T1

En esta etapa, se desarrolló un modelo de red neuronal recurrente tipo LSTM (Long Short-Term Memory) orientado a la clasificación de palabras habladas, utilizando como entrada los embeddings extraídos del modelo preentrenado Wav2Vec2.0, Wav2Vec2.0 permite representar señales acústicas como secuencias densas de vectores de alta dimensión (768), preservando la información fonética relevante mediante un preentrenamiento no supervisado sobre grandes corpus de voz. Los embeddings generados para cada grabación tienen forma (T, 768), donde T representa el número de pasos temporales resultante de la segmentación interna del modelo Wav2Vec2.0 (aproximadamente 20 ms por paso). Para poder procesar estas secuencias con redes LSTM, se ajustó la longitud de cada muestra a un valor fijo mediante padding (maxlen = 95), lo que permitió construir un tensor de entrada uniforme de forma (n_samples, 95, 768).

La entrada del modelo recibe secuencias de 95 pasos temporales, cada uno con 768 dimensiones, correspondientes a los embeddings de Wav2Vec2.0 y cuenta con tres capas LSTM bidireccionales cada una con 256 unidades por dirección. Las dos primeras capas son bidireccionales y permite que el modelo capture dependencias tanto hacia adelante como hacia atrás en la secuencia, lo cual es especialmente útil en tareas fonéticas para conservar la estructura temporal completa, mientras que la última colapsa la secuencia mediante, generando un vector resumen.

Por último, una capa densa de salida con activación softmax [17], esta capa convierte el vector de estado oculto final en una distribución de probabilidad sobre las clases (palabras objetivo). El número de salidas coincide con la cantidad de clases del problema.

Después de cada capa LSTM se aplicó una tasa de dropout 0.1032, con el fin de prevenir el sobreajuste sin afectar significativamente la capacidad del modelo para aprender patrones representativos. Además de Dropout, se incorporó regularización L2 1.07×10^{-6} en cada capa entrenable, penalizando los pesos excesivamente grandes para promover la generalización. El modelo fue optimizado mediante el algoritmo Adam, con una tasa de aprendizaje ajustada 3.77×10^{-4}, y entrenado con un tamaño de batch de 64 muestras.

Optimización con Optuna modelo LSTM_T2

La etapa de optimización es fundamental para determinar la configuración óptima de la LSTM en varios aspectos. El modelo cuenta con tres capas LSTM bidireccionales cada una con 192 unidades bidireccionales. Por último, una capa densa de salida con activación softmax, esta capa convierte el vector de estado oculto final en una distribución de probabilidad sobre las clases. Después de cada capa LSTM se aplicó una tasa de dropout 0.2352, con el fin de prevenir el sobreajuste sin afectar significativamente la capacidad del modelo para aprender patrones representativos. Además de dropout, se incorporó regularización L2 0.00023 en cada capa, penalizando los pesos excesivamente grandes para promover la generalización y entrenado con un

tamaño de batch de 64 muestras.

Optimización con Optuna modelo GRU

Se construyó un modelo secuencial basado en redes neuronales recurrentes del tipo GRU, Los mejores resultados fueron obtenidos con una arquitectura compuesta por cuatro capas GRU bidireccionales, cada una con 112 unidades por dirección. Cada capa GRU incluye una función bidireccional, salvo en la última, lo que permite capturar y transferir información contextual a lo largo de la secuencia antes de condensarla en un vector resumen. Este vector se conecta a una capa Dense con activación softmax, que produce la probabilidad para cada clase objetivo.

Para mejorar la capacidad de generalización del modelo y reducir el riesgo de sobreajuste, se implementaron las siguientes estrategias de regularización: dropout se aplicó una tasa de desactivación del 27.4% después de cada capa GRU, permitiendo que la red aprenda representaciones más robustas sin depender de conexiones específicas. Regularización L2 con un valor de penalización 1.52×10^{-6}, esta técnica ayudó a controlar el crecimiento excesivo de los pesos, fomentando soluciones más estables. El modelo fue optimizado con una tasa de aprendizaje de 1.21×10^{-3}, y entrenado en batches de 64 muestras durante un máximo de 50 épocas.

3.6 Indicador Visual y realimentación al usuario

Después de procesar y clasificar la señal de voz, el sistema genera un indicador visual a la señal de voz ingresada en la etapa 1, lo que implica la creación de representaciones visuales que reflejen las características y patrones identificados en la señal de voz. Además, se visualiza un porcentaje de similitud entre la palabra de referencia y la pronunciada por el usuario, y así poder generar una señal vibrotáctil equivalente a la palabra pronunciada.

3.6.1 Señales vibrotáctiles

La señal de entrada es una grabación de audio muestreada a 16 kHz y segmentada en ventanas de 25 ms con desplazamiento de 10 ms. Sobre estas ventanas se aplicó un banco de 32 filtros Gammatone, abarcando un rango de frecuencias desde los 150 Hz en adelante. La función espectrograma basado en filtros Gammatone (gtgram) se utilizó para generar un espectrograma gammatone, el cual proporciona una representación del contenido energético por banda auditiva a lo largo del tiempo. Cada fila del espectrograma representa una banda de frecuencia, y cada columna un instante temporal. La matriz resultante fue transpuesta para reorganizar los datos en un formato donde cada columna representa la evolución temporal de una banda de frecuencia específica.

Para reconstruir una señal temporal continua, se usa la transformada Hilbert y se interpola la envolvente de cada banda al tamaño completo de la señal original, utilizando interpolación lineal sobre un eje temporal regular como se observa en Figura 3.13.

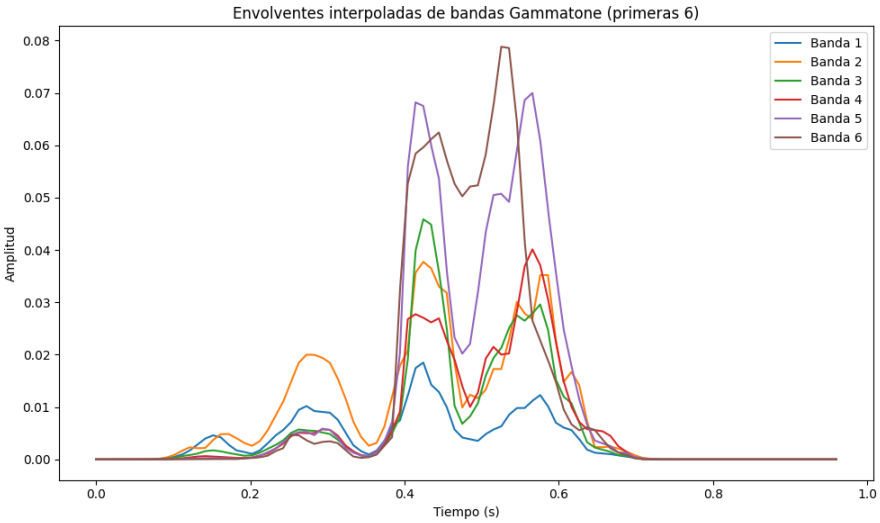

Figura 3.13 Primeras seis envolventes interpoladas de bandas Gammatone.

Fuente: Elaboración propia.

Posteriormente, se sumaron todas las bandas para obtener una única envolvente multimodal compuesta como se evidencia en la Figura 3.14, que representa la modulación de energía total percibida en el sistema auditivo simulado.

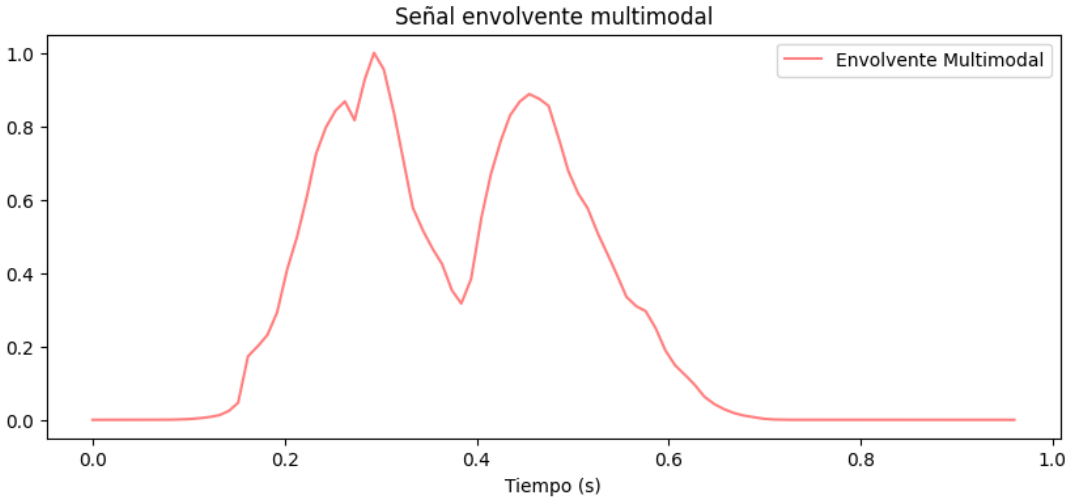

Figura 3.14 Envolvente multimodal de la señal de audio.

Fuente: Elaboración propia.

La envolvente multimodal fue normalizada entre 0 y 1 para garantizar un rango de amplitud constante. A continuación, se generó una portadora senoidal de 150 Hz, frecuencia elegida por su alta sensibilidad al tacto. La señal final fue obtenida mediante modulación de amplitud (AM) de la portadora por la envolvente multibanda. Finalmente, la señal fue normalizada en el rango [−1,1] para asegurar compatibilidad con sistemas de reproducción de audio vibrotáctil. La Figura 3.15 ilustra la señal vibrotáctil resultante.

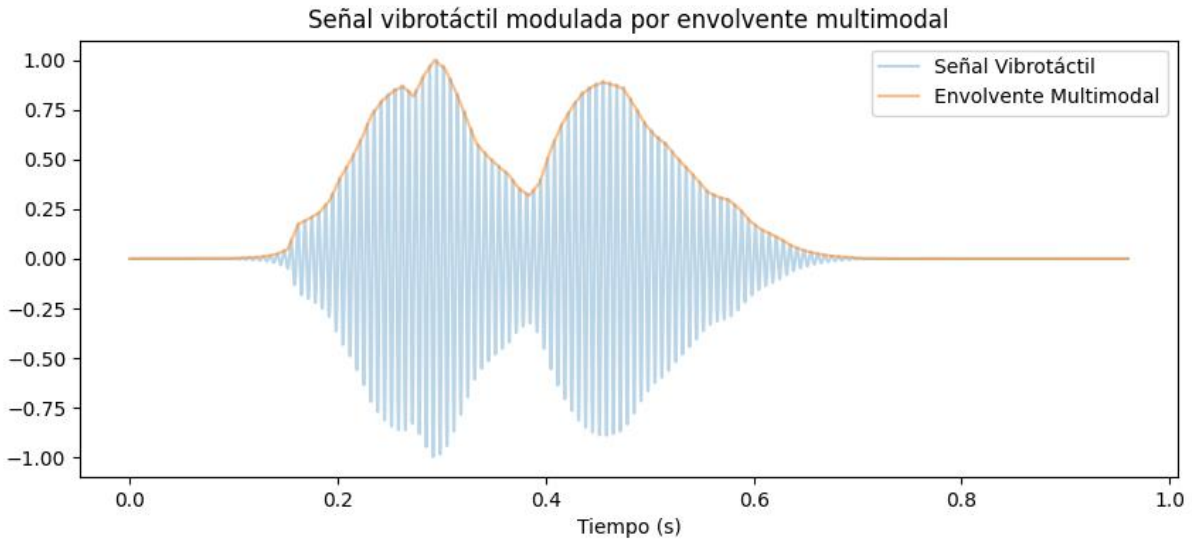

Figura 3.15 Señal vibro táctil por envolvente multimodal.

Fuente: Elaboración propia.

3.6.2 Similitud entre audios

Para este proceso se implementó un procedimiento basado en la comparación de descriptores acústicos relevantes extraídas de cada señal [54]. Este enfoque considera descriptores de baja frecuencia relacionados con la energía y el tono, como descriptores espectrales de mayor relevancia fonética como los formantes. Para cada una de las señales de audio analizadas, se calcularon cuatro descriptores fundamentales:

- Energía: se estimó como la suma de los cuadrados de las amplitudes de la señal en el dominio temporal, representando el contenido energético global de cada grabación.

- Pitch (Tono fundamental): se extrajeron los valores de frecuencia fundamental detectados en la señal, y se calculó su promedio sobre los valores positivos para obtener una medida robusta del tono dominante.

- Formantes (F1 y F2): se obtuvo la representación de formantes de cada señal, extrayendo los dos primeros formantes (F1 y F2), que se asocian con la configuración articulatoria de las vocales y son altamente discriminativos en tareas de reconocimiento del habla.

Debido a que los descriptores extraídos tienen escalas y unidades distintas, se aplicó una normalización tipo Z-score para cada vector de características. De este modo, se asegura que todas las características estén en una misma escala, centradas en cero con varianza 1, lo que permite una comparación equitativa entre dimensiones como se puede ver en la Figura 3.16.

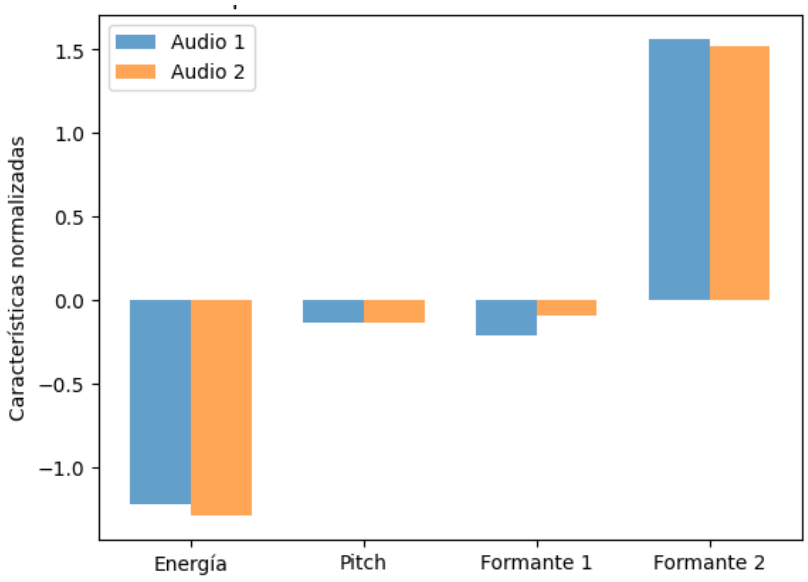

Figura 3.16 Descriptores normalizados.

Fuente: Elaboración propia.

Para calcular el grado de similitud entre dos vectores normalizados, se utilizó la distancia euclidiana ponderada, En este caso, se otorgó mayor relevancia a los formantes (F1 y F2)[54] mediante una ponderación de [0.1,0.1,0.4,0.4] [54], considerando su importancia en la caracterización articulatoria. A partir de la distancia obtenida, se estimó un porcentaje de similitud mediante una transformación inversa que considera la distancia máxima posible bajo las mismas condiciones de normalización y ponderación.

3.6.3 Interfaz de usuario

La interfaz gráfica fue desarrollada utilizando el framework PyQt5 con el fin de proporcionar una herramienta interactiva e intuitiva para la gestión, grabación, procesamiento y análisis de señales de voz, orientada a tareas de reconocimiento y comparación acústica. La interfaz está diseñada para facilitar el trabajo con un conjunto predefinido de palabras organizadas por fonemas, y ofrece funcionalidades claves como visualización jerárquica de grabaciones, generación de señales vibrotáctiles, reconocimiento automático del habla y evaluación de similitud acústica.

Registro e identificación del usuario

El usuario inicia el proceso registrando su información personal (nombre, apellido, edad y género) a través de campos habilitados en la parte superior izquierda de la interfaz. Esta información es utilizada para generar una carpeta única en el sistema de archivos, bajo la estructura: Usuarios/Nombre Apellido Edad Género.

Cada vez que se realiza una grabación, esta se almacena dentro de una subcarpeta que refleja la organización jerárquica por fonema y palabra (por ejemplo: Usuarios/Juan

Pérez 25 Masculino/Palatales/calle/) además almacenando la fecha de cada grabación para analizar el historial, guardando en una base de datos toda la información del usuario. Esta estructura permite mantener un orden lógico y segmentado de las muestras, facilitando tanto la trazabilidad como el análisis posterior.

Grabación y almacenamiento de muestras de voz

Al presionar el botón de grabación, el sistema inicia la captura de la señal de audio a través del micrófono con una frecuencia de muestreo fija de 16 kHz y a un solo canal. La grabación se detiene al presionar nuevamente el mismo botón. Antes de ser almacenada, la señal pasa por un proceso de preprocesamiento, que incluye:

- Filtrado pasa altos para eliminar componentes de baja frecuencia no relevantes.
- Reducción de ruido utilizando una muestra del primer segundo como estimación del ruido ambiente.
- Normalización y conversión a formato .wav con frecuencia de muestreo de 16 kHz.

Si existe ya una grabación con el mismo nombre, el sistema evita sobrescribirla, añadiendo automáticamente un sufijo numérico incremental al nombre del archivo.

Navegación y visualización de grabaciones

La interfaz permite explorar las grabaciones almacenadas de forma jerárquica. Las grabaciones se agrupan por fonema y palabra dentro de la carpeta del usuario seleccionado. Al hacer doble clic sobre un archivo .wav, el sistema carga la señal correspondiente, la gráfica y actualiza la variable interna de selección para su posterior análisis o reproducción. Adicionalmente, mediante un clic derecho sobre un archivo, se despliega un menú contextual que permite eliminar grabaciones de forma segura, previa confirmación del usuario.

Reproducción y generación de señales vibrotáctiles

La interfaz ofrece la posibilidad de reproducir tanto la última grabación realizada como cualquier grabación seleccionada desde el árbol de exploración. Además, incorpora un módulo para la generación de señales vibrotáctiles basadas en una transformación de Gammatone multibanda y modulación de amplitud (AM) con una portadora de 150 Hz. Esta señal se calcula a partir de la transformada Hilbert y envolvente multibanda interpolada y se reproduce como una señal continua de estimulación, siendo además graficada para su visualización. Las señales vibrotáctiles fueron transmitidas mediante un par de parlantes J&R J5219. Estos dispositivos cuentan con una membrana de silicona lateral, sobre la cual los participantes apoyaban la palma de la mano para percibir la vibración generada por la señal acústica. La selección de este modelo respondió a su capacidad de reproducir bajas frecuencias, generando una respuesta táctil perceptible en la membrana.

Comparación de similitud acústica

El sistema permite comparar cualquier grabación del usuario con una referencia almacenada en la carpeta Referencias/<Palabra>/<Género>/audio.wav. Para esta comparación, se extraen los siguientes descriptores acústicos: Energía total, Frecuencia fundamental (pitch), Primer y segundo formante (F1 y F2).

Ambas señales (usuario y referencia) se normalizan mediante z-score, y se calcula una distancia euclidiana ponderada que es convertida en porcentaje de similitud. El resultado se presenta mediante una barra de progreso coloreada dinámicamente (rojo, amarillo o verde) y un mensaje descriptivo que indica si la similitud es alta, aceptable o baja.

Esta interfaz integra de forma eficiente los módulos de grabación, análisis, comparación y visualización de señales de voz, ofreciendo una herramienta potente y accesible para tareas de investigación en reconocimiento de voz, estimulación vibrotáctil y procesamiento acústico personalizado. Su diseño modular permite una fácil extensión para futuras funcionalidades como clasificación automática, entrenamiento supervisado y retroalimentación interactiva al usuario.

Figura 3.17 Interfaz de usuario.

Fuente: Elaboración propia.

3.7 Resumen

La solución propuesta aborda las tareas de grabar y reproducir de señales de voz,

procesamiento de señales de voz, clasificación automática de palabras habladas mediante un sistema que integra dos enfoques complementarios para la extracción de características acústicas: uno basado en característica predefinidas y otro sustentado en aprendizaje profundo mediante modelos preentrenados. Ambos enfoques alimentan modelos secuenciales (LSTM y GRU), optimizados para maximizar el rendimiento del sistema en términos de precisión y generalización, visualización de descriptores acústicos y reproducción de señales vibrotáctiles.

La etapa de adquisición de la señal consistió en la grabación de muestras de voz en un, utilizando una frecuencia de muestreo de 16 kHz en configuración monocanal. Posteriormente, se llevó a cabo un proceso de preprocesamiento de las señales para mejorar la calidad y facilitar la extracción de características relevantes. Inicialmente, se aplicó un filtro pasa altas con una frecuencia de corte de 120 Hz, con el objetivo de eliminar componentes de baja frecuencia no asociadas a la voz. Seguidamente, se implementó una técnica de reducción de ruido mediante sustracción espectral, la cual permite estimar y sustraer el espectro del ruido ambiente, incrementando así la relación señal-ruido sin comprometer la inteligibilidad de la señal.

El primer enfoque parte de la extracción de características espectrales utilizando coeficientes cepstrales en las frecuencias de Mel (MFCC), junto con sus primeras y segundas derivadas (delta y delta-delta), el espectrograma y las características chroma. Cada señal de voz es segmentada en ventanas de 25 ms con un solapamiento del 50%, y se calcula un vector de características de 39 dimensiones por frame. Posteriormente, se aplica padding temporal para igualar la longitud de las secuencias y se codifican las etiquetas. Estas representaciones sirven como entrada para modelos secuenciales LSTM y GRU, cuya arquitectura se compone de dos a cuatro capas recurrentes con funciones de regularización como Dropout y L2. La salida de estas redes se conecta a una capa densa con activación softmax para la predicción multiclase. Los hiperparámetros del modelo, incluyendo el número de unidades, tasa de abandono, tipo de RNN (LSTM o GRU), regularización L2, tasa de aprendizaje y tamaño de batch, son ajustados mediante Optuna, maximizando la métrica precisión sobre un conjunto de validación.

El segundo enfoque hace uso del modelo preentrenado Wav2Vec2.0 (facebook/wav2vec2-base) como extractor automático de características profundas. Cada grabación es normalizada y remuestreada a 16 kHz, y luego procesada por el modelo Wav2Vec2.0 para obtener una secuencia de vectores de 768 dimensiones, que encapsulan información fonética y temporal de alto nivel. Estas secuencias son truncadas o rellenadas hasta un valor fijo (maxlen = 95) para permitir el procesamiento por redes recurrentes. Al igual que en el enfoque manual, se utilizan modelos LSTM y GRU con arquitectura secuencial, optimizados con búsqueda de hiperparámetros. La combinación de embeddings ricos y arquitectura recurrente permite capturar relaciones temporales complejas y obtener modelos altamente discriminativos para la clasificación de las 12 palabras habladas.

Ambos enfoques comparten una estrategia común de evaluación y validación cruzada, así como el uso de métricas macro como precisión, recall y F1-Score para asegurar un

desempeño equilibrado entre clases. Se observa que, dependiendo de la configuración del modelo y el tipo de representación usada, las redes GRU pueden superar a las LSTM en términos de eficiencia y rendimiento, especialmente al utilizar embeddings de Wav2Vec2.0. Esto sugiere que la combinación de características extraídas automáticamente con arquitecturas recurrentes ligeras es una alternativa eficaz para el reconocimiento de voz con recursos computacionales moderados.

El procedimiento de generación de señales vibrotáctiles se basa en simular el procesamiento auditivo humano mediante un banco de filtros Gammatone, que descompone una señal de audio en múltiples bandas de frecuencia. A cada banda se aplica la transformada Hilbert y se extrae su envolvente temporal, y estas envolventes se interpolan y suman para formar una envolvente compuesta que representa la energía global del estímulo auditivo. Esta envolvente modula la amplitud de una portadora senoidal de 150 Hz, frecuencia óptima para la percepción táctil, generando así una señal final que puede ser reproducida por un parlante. El resultado es una vibración cuya intensidad varía en el tiempo de forma análoga a la estructura rítmica y espectral del habla, permitiendo que el usuario perciba y discrimine palabras a través del sentido del tacto.

En la etapa de medición de similitud entre audios se realiza el cálculo y comparación de descriptores acústicos: energía, pitch (tono fundamental) y los dos primeros formantes (F1 y F2). Estas características fueron normalizadas mediante la técnica de Z-score para garantizar una comparación en escalas equivalentes. Posteriormente, se aplicó una distancia euclidiana ponderada, asignando mayor peso a los formantes por su importancia fonética. Finalmente, la distancia obtenida se transformó en un porcentaje de similitud, donde valores cercanos al 100% indican alta correspondencia entre los audios, permitiendo una evaluación precisa y ajustable del grado de similitud acústica.

4. Pruebas y Resultados

En este capítulo se presenta los resultados de las pruebas realizadas para los métodos propuestos en el Capítulo 3, para la clasificación de voz y la generación de señales vibro táctiles. Se describe y analizan los resultados obtenidos de los modelos ocultos de Markov (HMM), LSTM y GRU, se comparan en términos de Precision, Recall y F1-Score. Finalmente se realiza un análisis comparativo del desempeño de los modelos con la literatura, destacando sus ventajas y limitaciones.

4.1. Pruebas y resultados de los modelos de aprendizaje

Se realizaron múltiples pruebas para evaluar el rendimiento de los modelos HMM, LSTM y GRU en la tarea de clasificación de 12 palabras habladas. Se evaluaron dos enfoques de extracción de características: (1) extracción de MFCCs y derivadas deltas, espectrograma y características chorma, y (2) uso de embeddings generados por el modelo preentrenado Wav2Vec2.0. En ambos casos, los modelos fueron entrenados mediante validación cruzada estratificada (k-fold=5) y optimizados utilizando la biblioteca Optuna con búsqueda de hiperparámetros.

Las métricas utilizadas para la evaluación fueron:

- Precisión: media de la precisión individual por clase.
- Recall: media de la sensibilidad por clase.
- F1-score macro: promedio armónico entre la precisión y la sensibilidad, útil en escenarios con desbalance de clases.

Se usaron en total 745 grabaciones entre todas las 12 palabras, para cada clase se dividieron en 70% para entrenamiento y 30% para validación, separando por usuarios, así evitando tener grabaciones del mismo usuario en los dos conjuntos de datos.

4.1.1. Pruebas y resultados para modelos con extracción de características predefinidas

Se implementó dos máquinas de aprendizaje: HMM (sección 3.5.3.1) y LSTM (sección 3.5.3.2).

Modelos ocultos de Márkov

Se emplearon Modelos Ocultos de Markov (HMM) multivariantes con observaciones continuas. Para cada clase, se entrenó un modelo HMM con un número específico de estados ocultos, para las palabras “Bicicleta” y “Tarjeta”, se asignaron 5 y 6 estados para cada género, mientras que para las otras clases se utilizó un valor por defecto de 4 componentes. El número de iteraciones del algoritmo de entrenamiento (Baum-Welch) se fijó en 100.

Cada modelo fue inicializado con una matriz de transición uniforme entre estados y parámetros aleatorios para las distribuciones gaussianas diagonales. Para evitar el sobreajuste y promover una distribución más suave, se aplicó una regularización de tipo L2 a las probabilidades iniciales "*startprob*" y de transición *"transmat"*. Esta regularización consistió en una multiplicación escalar con un factor de penalización (0.01), seguida de una normalización para asegurar que las distribuciones resultantes sumaran 1, garantizando así la validez probabilística del modelo.

Modelos LSTM1 y LSTM2

Para los modelos se utilizó el optimizador *"Adam"*, la función de perdida de entropía cruzada. Además, para mejorar la eficiencia del entrenamiento se aplica la técnica de "Early Stopping", monitoreando la perdida en el conjunto de validación y detenido el entrenamiento si no se observa mejora después de 5 épocas consecutivas. En la Figura 4.1, se pueden apreciar los resultados del modelo LSTM1 a lo largo de 10 épocas y en la Figura 4.2, se observa los resultados para el modelo LSTM2 a lo largo de 17 épocas y sus matrices de confusión.

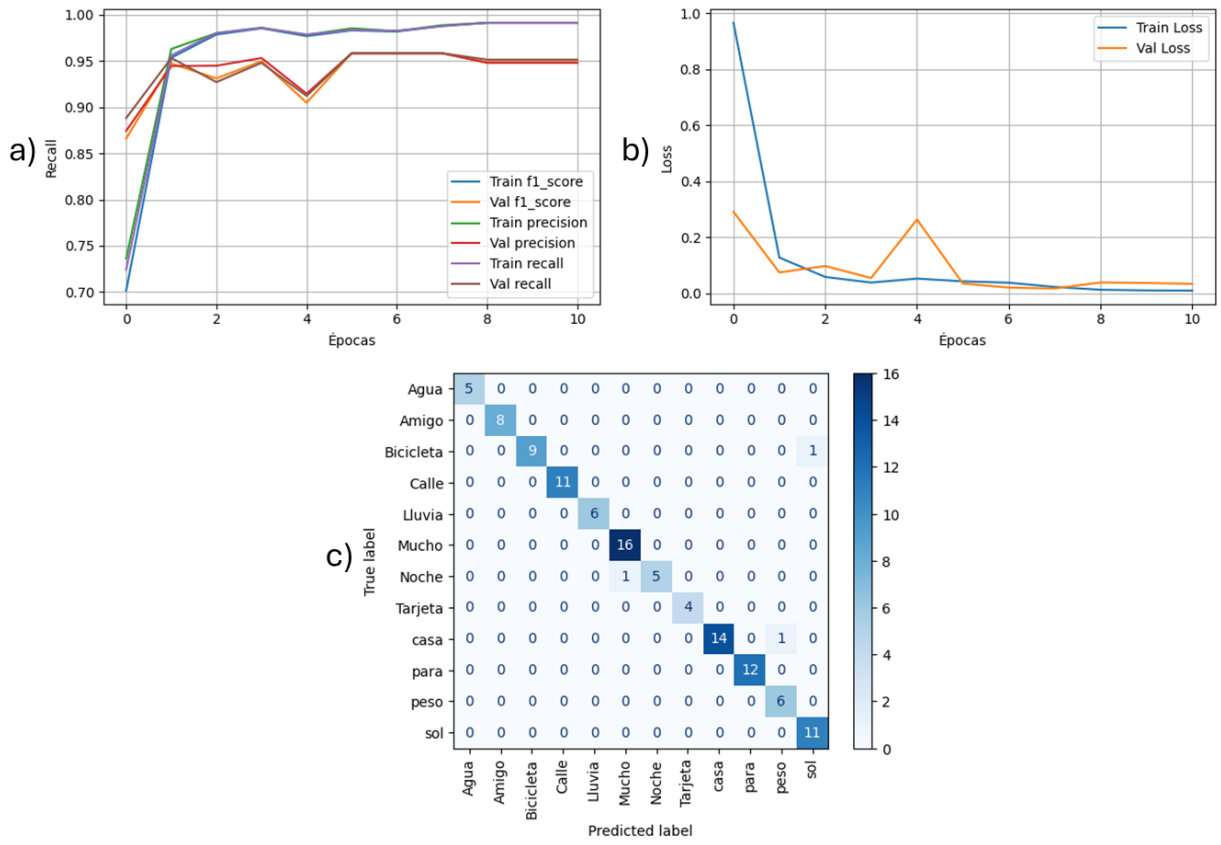

Figura 4.1 Modelo LSTM1, a) Métricas por épocas, b) Perdida por épocas, c) Matriz de confusión.

Fuente: Elaboración propia.

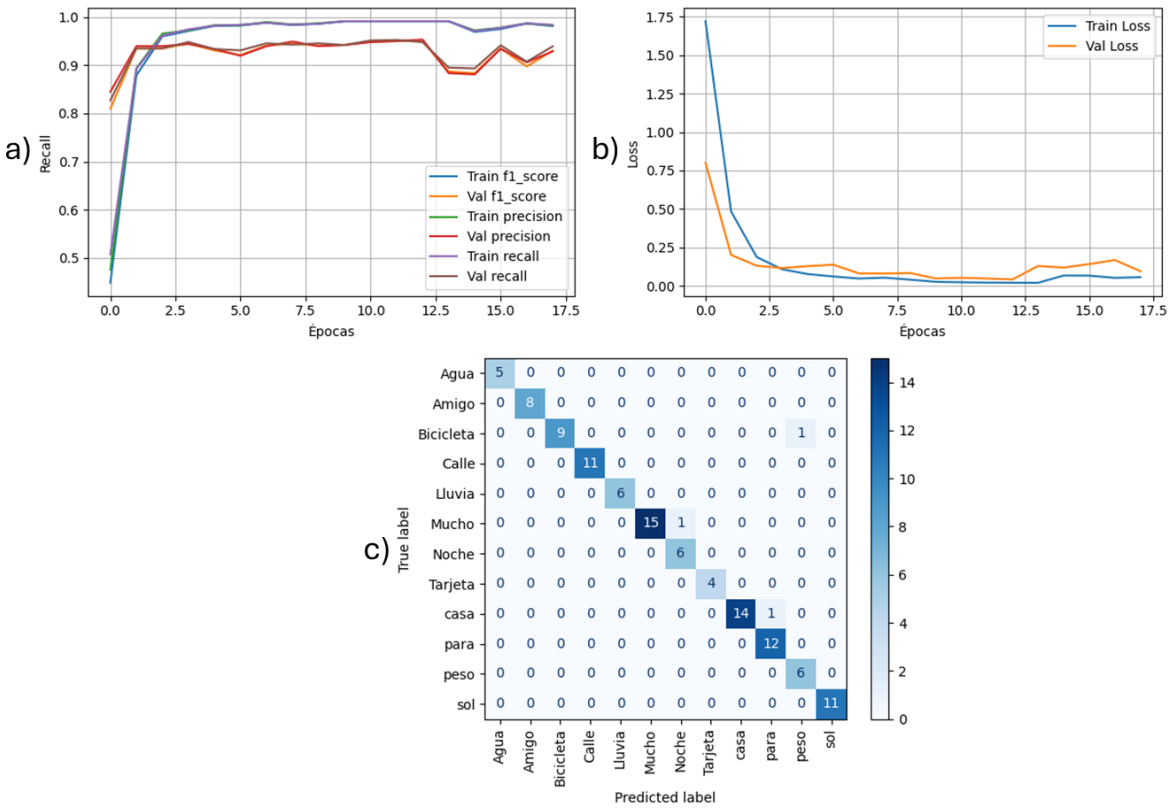

Figura 4.2 Modelo LSTM2, a) Métricas por épocas, b) Perdida por épocas, c) Matriz de confusión.

Fuente: Elaboración propia.

En la Figura 4.2 se observa cómo las curvas de precisión convergen, indicando que el modelo mejora su capacidad de generalización a medida que avanza el entrenamiento. Las curvas de costo (Loss) muestran una tendencia decreciente cercana a 0 tanto en el conjunto de entrenamiento como en el de validación.

Adicionalmente se realizó el cálculo de las métricas Recall, F1-Score y Precisión anexadas en la Tabla 4.1 Resultados modelos.Tabla 4.1.

Tabla 4.1 Resultados modelos.

Fuente: Elaboración propia.

Resultados para entrenamiento					Resultados para validación			
Modelo	**Recall**	**F1-Score**	**Precision**	**Loss**	**Recall**	**F1-Score**	**Precision**	**Loss**
HMM	0.85±0.06	0.84±0.05	0.84±0.06		0.79	0.79	0.79	
LSTM1	0.99±0.02	0.99±0.02	0.99±0.02	0.01	0.95	0.94	0.94	0.04
LSTM2	0.99±0.01	0.98±0.02	0.99±0.01	0.06	0.95	0.95	0.95	0.09

4.1.2 Prueba y resultados con embeddings de Wav2Vec2.0

Se implementó dos máquinas de aprendizaje LSTM y GRU (sección 3.6), Para los modelos se utilizó el optimizador *"Adam"*, la función de perdida de entropía cruzada. Además, para mejorar la eficiencia del entrenamiento se aplica la técnica de *"Early Stopping"*, monitoreando la perdida en el conjunto de validación y detenido el entrenamiento si no se observa mejora después de 5 épocas consecutivas y la técnica de *"Reduce LearningRate"*, reduciendo la tasa de aprendizaje si en 3 épocas la validación no ha sufrido cambios. En la Figura 4.3, se pueden apreciar los resultados del modelo LSTM_T1 a lo largo de 20 épocas. En la Figura 4.4, se observa los resultados para el modelo LSTM_T2 a lo largo de 20 épocas. En la Figura 4.5, se observa los resultados del modelo GRU a lo largo de 17 épocas.

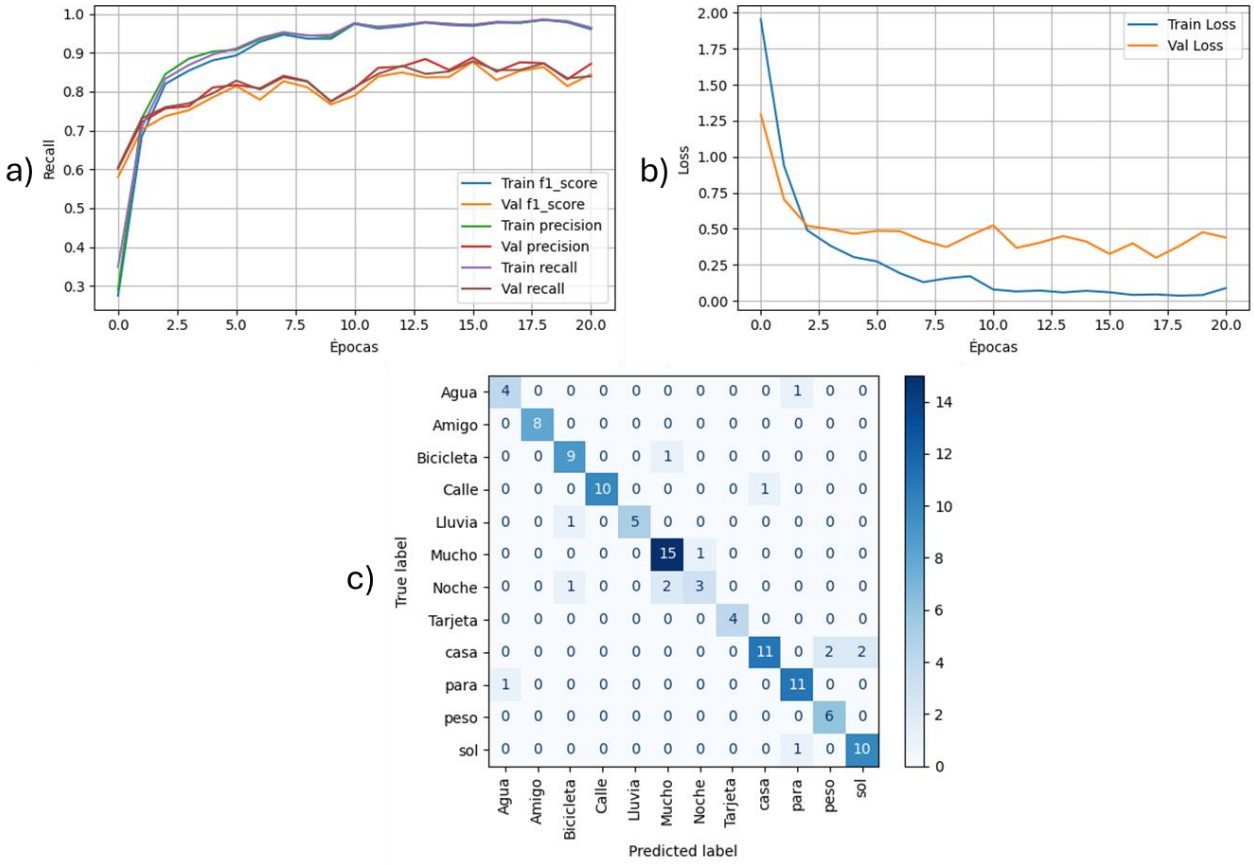

Figura 4.3 Modelo LSTM_1, a) Gráfica de métricas b) Gráfica de perdida, c) Matriz de confusión.

Fuente: Elaboración propia.

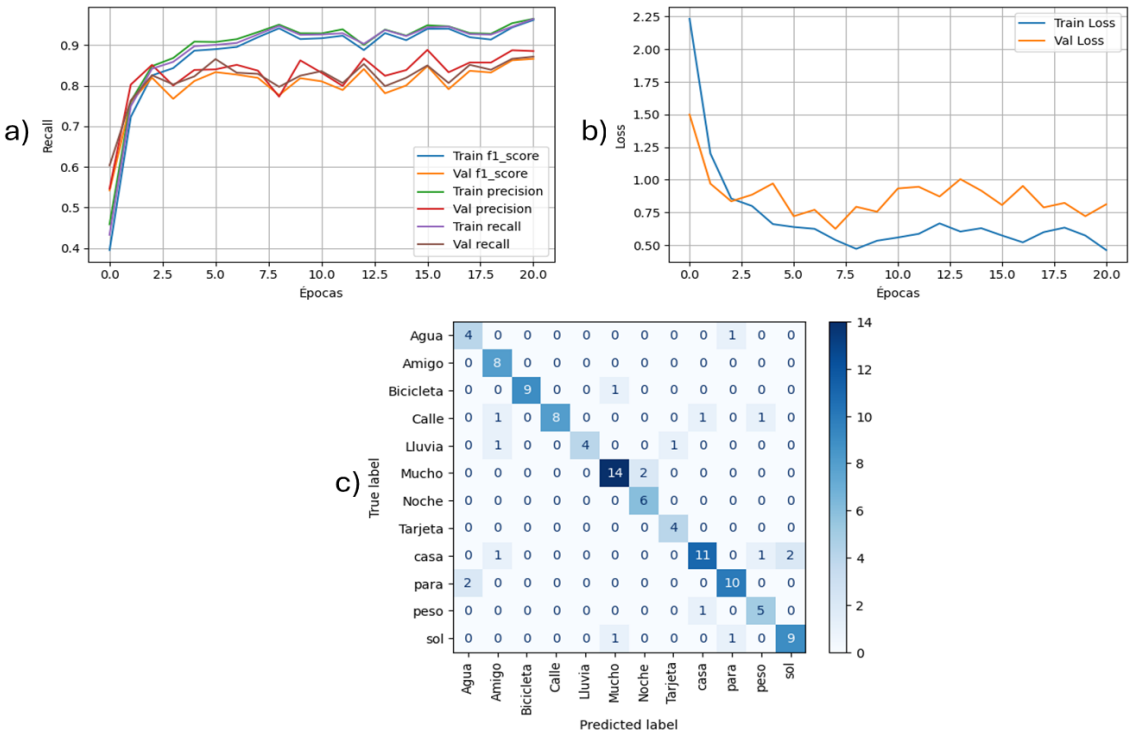

Figura 4.4 Modelos LSTM_T2 a) Gráfica de métricas b) Gráfica de perdida, c) Matriz de confusión.

Fuente: Elaboración propia.

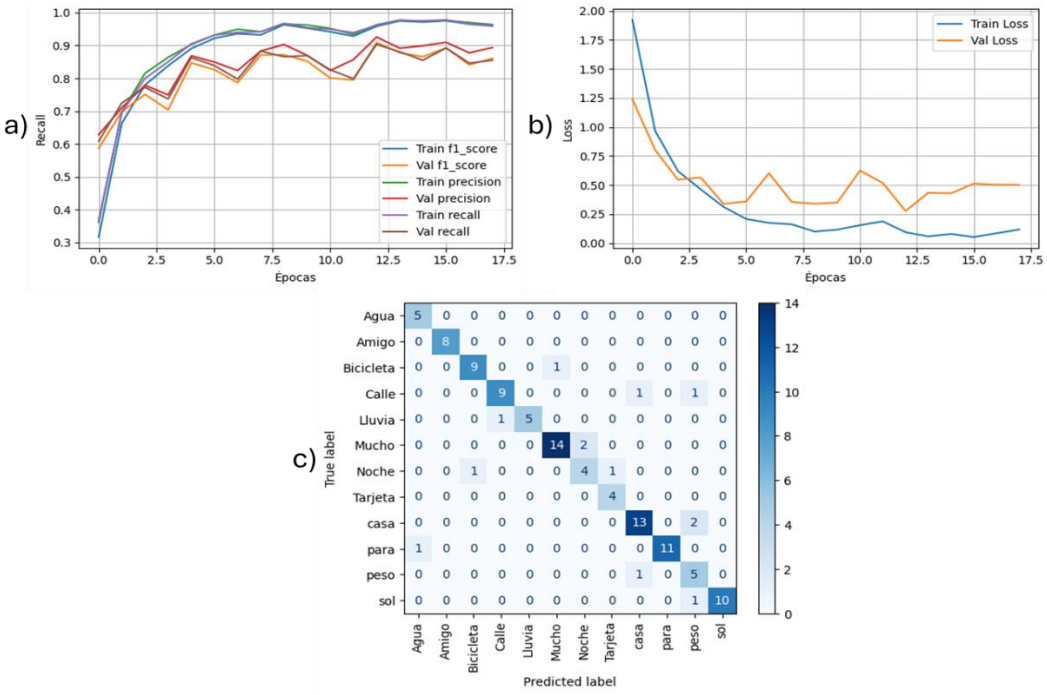

Figura 4.5 Modelo GRU, a) Gráfica de métricas b) Gráfica de perdida, c) Matriz de confusión.

Fuente: Elaboración propia.

En la Figura 4.5 se observa cómo las curvas de precisión convergen, indicando que el modelo mejora su capacidad de generalización a medida que avanza el entrenamiento. Las curvas de costo (Loss) muestran una tendencia decreciente cercana a 0 en el conjunto de entrenamiento, pero en el de validación la perdida entre 0.3 y 0.7. Adicionalmente se realizó el cálculo de las métricas Recall, F1-Score y Precisión anexadas en la Tabla 4.2.

Tabla 4.2 Resultados modelos con embeddings de Wav2Vec2.0.

Fuente: Elaboración propia.

Resultados para entrenamiento					Resultados para validación			
Modelo	**Recall**	**F1-Score**	**Precisión**	**Loss**	**Recall**	**F1-Score**	**Precisión**	**Loss**
LSTMT1	0.98±0.03	0.98±0.03	0.99±0.04	0.03	0.84	0.84	0.87	0.44
LSTMT2	0.96±0.02	0.96±0.02	0.97+0.02	0.47	0.87	0.87	0.88	0.81
GRU	0.96±0.02	0.96±0.02	0.960.01	0.12	0.85	0.86	0.89	0.50

4.1.3 Análisis comparativo para los modelos de aprendizaje

Se evaluaron diferentes combinaciones de métodos de extracción de características y arquitecturas de modelos de aprendizaje secuencial con el fin de determinar cuál estrategia ofrecía un mejor desempeño en la tarea de clasificación de palabras habladas. Para ello, se implementaron dos líneas metodológicas principales: una basada en la extracción de características predefinidas (MFCC, deltas, espectrograma y chroma) y otra sustentada en la generación de embeddings acústicos automáticos mediante el modelo preentrenado Wav2Vec2.0. Cada conjunto de características fue evaluado sobre distintas arquitecturas: HMM, LSTM y GRU, con ajuste de hiperparámetros y técnicas de regularización que promovieran la generalización y eficiencia.

En primer lugar, los resultados con características predefinidas mostraron que el modelo HMM multivariante logró un rendimiento razonable, con una F1-score de validación de 0.7906. Si bien este enfoque modela de manera explícita las transiciones de estados en el tiempo, sus supuestos probabilísticos y su limitada capacidad de representación lo ubicaron por debajo de las arquitecturas basadas en aprendizaje profundo.

En contraste, los modelos LSTM entrenados con las mismas características predefinidas obtuvieron mejores resultados. El modelo LSTM1 alcanzó un F1-score de validación de 0.9486, con una pérdida baja de 0.0395, evidenciando una capacidad notable para capturar patrones temporales en las secuencias acústicas. El modelo LSTM2, aunque mostró un rendimiento ligeramente inferior (val_F1 = 0.95), presentó una validación más estable durante más épocas, lo cual puede relacionarse con diferencias en la profundidad o en la configuración de regularización.

En una segunda fase, se evaluaron modelos entrenados con embeddings generados automáticamente por Wav2Vec2.0, lo que permitió utilizar representaciones acústicas más ricas y profundas. Estos embeddings fueron procesados por redes LSTM y GRU con varias capas recurrentes. El modelo LSTM_T1 logró un val_F1-score de 0.84, mientras que el modelo LSTM_T2 alcanzó un mejor desempeño con 0.87. Por su parte, el modelo GRU superó ligeramente a ambos modelos LSTM, obteniendo un F1-score de validación de 0.8607 y una precisión de 0.8936. Esta mejora, aunque ligera, puede atribuirse a la mayor eficiencia de las GRU para capturar dependencias temporales relevantes sin incurrir en la complejidad computacional que implican las LSTM.

Aunque los modelos con características predefinidas alcanzaron métricas más altas en validación, esto puede explicarse por un mayor grado de control en el preprocesamiento y un menor grado de variabilidad temporal. No obstante, los modelos basados en Wav2Vec2.0 muestran un desempeño competitivo, destacándose su capacidad de trabajar directamente sobre datos acústicos sin necesidad de diseñar manualmente las características. Este enfoque representa una solución escalable y flexible, especialmente adecuada para contextos donde la diversidad fonética es alta o no se cuenta con expertos en procesamiento de voz.

4.2 Pruebas y resultados del sistema para generar señales vibrotáctiles

Con el objetivo de evaluar el sistema de estimulación vibrotáctil, se diseñó y aplicó un protocolo experimental para evaluar la capacidad de reconocimiento táctil de palabras transmitidas mediante vibraciones generadas por un sistema de estimulación vibrotáctil basado en parlantes. Participaron seis usuarios con audición normal (U1, U2, U3, U4, U5 y U6), quienes fueron expuestos a una lista de 12 palabras agrupadas según su lugar de articulación fonética en español: alveolares ("casa", "para", "sol"), bilabiales ("mucho", "peso", "bicicleta"), palatales ("lluvia", "calle", "noche") y velares ("agua", "amigo", "tarjeta").

Para reducir la fatiga de los participantes, las sesiones se distribuyeron en cuatro días consecutivos. En cada jornada se trabajó con un conjunto de cuatro palabras, seleccionadas de modo que representaran diferentes tipos fonéticos. El protocolo incluyó dos fases:

Fase de entrenamiento: se reprodujo la señal vibrotáctil de cada palabra durante 1 minuto, con un total de cinco repeticiones por palabra. Durante esta fase, los participantes usaron audífonos con ruido blanco de fondo para evitar la percepción acústica directa del estímulo. El bloque completo tuvo una duración aproximada de 20 minutos por sesión.

Fase de prueba: se presentaron cuatro estímulos por palabra en un orden aleatorio. En cada ensayo, el participante debía identificar cuál de las cuatro palabras había sido transmitida a través de la vibración recibida. Las respuestas se registraron en un formato binario, asignando 1 a los aciertos y 0 a los errores.

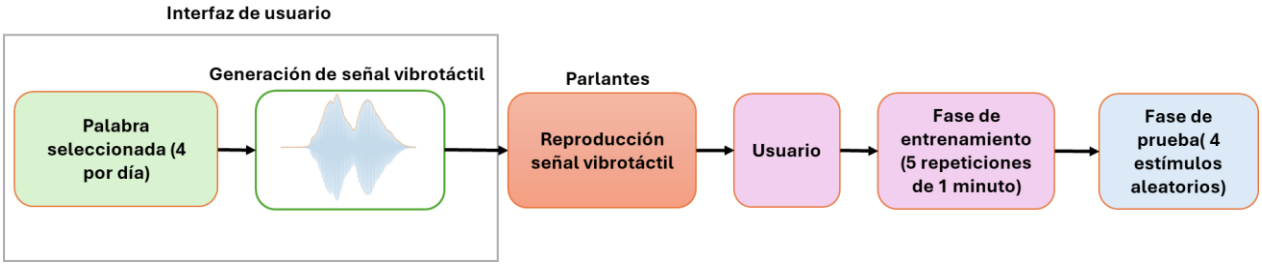

Figura 4.6 Diagrama de bloques protocolo de prueba del sistema de generación de señales vibrotáctiles.

Fuente: Elaboración propia.

Se recolectaron los resultados en la Tabla 4.3, donde cada celda indica si el usuario identificó correctamente el estímulo (valor 1) o se equivocó (valor 0).

Tabla 4.3 Tabla de resultados para el protocolo de prueba de generación de señales vibrotáctiles.

Fuente: Elaboración propia.

	CASA							CALLE							AGUA							BICICLETA					
	U1	U2	U3	U4	U5	U6		U1	U2	U3	U4	U5	U6		U1	U2	U3	U4	U5	U6		U1	U2	U3	U4	U5	U6
A	1	1	1	0	1	0	A	1	0	1	0	1	0	A	1	1	0	0	1	1	A	1	1	1	1	1	1
B	0	1	1	1	1	1	B	0	1	1	1	1	1	B	0	1	1	1	1	0	B	0	1	0	1	1	0
C	1	1	1	1	0	1	C	0	1	0	1	1	0	C	1	1	1	1	1	1	C	1	1	1	1	1	1
D	1	1	0	0	1	1	D	1	1	0	1	0	1	D	1	0	1	0	0	1	D	0	1	1	1	0	1

	PARA							MUCHO							LLUVIA							AMIGO					
	U1	U2	U3	U4	U5	U6		U1	U2	U3	U4	U5	U6		U1	U2	U3	U4	U5	U6		U1	U2	U3	U4	U5	U6
A	1	1	1	0	1	1	A	1	1	1	1	0	1	A	0	1	1	1	0	1	A	1	1	1	1	1	1
B	1	1	0	1	1	1	B	1	1	1	1	0	0	B	1	0	1	0	0	0	B	1	1	0	1	1	1
C	1	1	1	1	0	1	C	1	1	1	1	1	1	C	1	1	1	1	1	1	C	1	1	1	1	1	0
D	0	1	1	1	1	0	D	1	1	1	1	1	1	D	0	0	0	1	1	0	D	1	0	1	1	1	1

	SOL							PESO							NOCHE							TARJETA					
	U1	U2	U3	U4	U5	U6		U1	U2	U3	U4	U5	U6		U1	U2	U3	U4	U5	U6		U1	U2	U3	U4	U5	U6
A	1	1	0	1	1	0	A	1	1	0	1	0	1	A	1	1	1	1	1	1	A	1	1	0	1	0	1
B	1	1	1	0	1	1	B	1	1	1	1	0	1	B	1	1	1	1	1	1	B	0	1	0	0	0	1
C	1	1	1	1	1	1	C	0	1	0	1	1	1	C	1	1	1	1	1	0	C	0	1	0	1	1	0
D	0	1	1	1	0	1	D	0	0	1	0	1	1	D	1	1	1	1	0	0	D	1	0	1	1	1	1

Los resultados individuales por palabra y por usuario se consolidaron en la Tabla 4.3, donde se refleja el número de aciertos por palabra para cada uno de los cuatro intentos realizados por los seis participantes. A partir de esta matriz se calculó la precisión promedio por palabra, cuyos resultados se muestran en la Tabla 4.4.

Tabla 4.4 Precisión por palabra.

Fuente: Elaboración propia.

Palabra	Precisión	Palabra	precisión	Palabra	precisión
Mucho	0.88±0.07	Bicicleta	0.79±0.19	Peso	0.67±0.17
Amigo	0.88±0.06	Sol	0.79±0.10	Calle	0.62±0.14
Noche	0.84±0.08	Casa	0.75±0.16	Lluvia	0.58±0.20
Para	0.83±0.13	Agua	0.71±0.11	Tarjeta	0.58±0.17

Se observó un reconocimiento particularmente alto para las palabras amigo y mucho, con una precisión del 88% ±7% y ±6%. En contraste, las palabras tarjeta, calle, lluvia y peso presentaron las menores tasas de acierto, posiblemente debido a similitudes en sus firmas vibrotáctiles o a una menor discriminabilidad táctil. De la Tabla 4.3 y Tabla 4.4 se puede calcular la precisión para cada usuario para las 12 palabras.

Tabla 4.5 Precisión por usuario.

Fuente: Elaboración propia.

Usuario	Precisión
U1	0.71
U2	0.88
U3	0.71
U4	0.79
U5	0.69
U6	0.71

El usuario U2 presentó el mejor desempeño global con un 88% de aciertos, mientras que el usuario U5 tuvo el rendimiento más bajo, con un 69%. Esta variabilidad puede atribuirse a factores individuales como la sensibilidad táctil, la concentración o la familiaridad con la tarea.

4.2.1 Análisis para las pruebas y resultados de la generación de señales vibrotáctiles

El análisis de los resultados evidencia una tasa de reconocimiento global positiva, con una precisión promedio de 74.8% ±7% entre los seis usuarios. Las palabras amigo y mucho alcanzaron una precisión del 88%, lo cual sugiere que sus patrones vibrotáctiles presentan una estructura rítmica o dinámica claramente discriminable. Por el contrario, las palabras tarjeta 58% y calle 62% mostraron los niveles de reconocimiento más bajos, lo que indica posibles limitaciones del sistema para representar diferencias fonéticas finas o la presencia de firmas táctiles similares entre ciertos fonemas.

El desempeño por usuario también refleja una variabilidad interindividual, siendo el usuario U2 el de mayor precisión 88% y U5 el de menor 69%. Estas diferencias podrían

deberse a factores como la sensibilidad cutánea, la atención sostenida o la estrategia de aprendizaje empleada por cada participante. Los resultados confirman la viabilidad del sistema para transmitir información lingüística por vía táctil utilizando señales vibrotáctiles generadas con modulación de envolvente.

4.3 Discusión

Se implementaron dos enfoques para el reconocimiento de palabras a partir de señales acústicas preprocesadas: modelos ocultos de Markov (HMM) con características manuales y redes neuronales LSTM y GRU alimentadas con embeddings extraídos de Wav2Vec2.0. Los HMM multivariantes, entrenados con observaciones continuas, mostraron un rendimiento adecuado F1-score aproximado a 0.84 en validación, particularmente al incorporar regularización L2 para evitar sobreajuste. Sin embargo, los modelos LSTM con características manuales superaron ampliamente a los HMM val_F1=0.95, lo que indica una mayor capacidad de modelado de secuencias temporales complejas. Esta tendencia se mantuvo al incorporar representaciones profundas de audio: aunque los LSTM y GRU mostraron alto rendimiento en entrenamiento, el desempeño en validación fue más modesto val_F1=0.86, reflejando un posible desajuste entre las características aprendidas y la variabilidad en los datos de prueba. Este comportamiento sugiere que los embeddings de Wav2Vec2.0, aunque potentes, requieren calibración cuidadosa del modelo de clasificación para evitar sobreajuste y mejorar la generalización.

Complementariamente, se diseñó un protocolo de evaluación sensorial para medir la precisión con que los usuarios identificaban palabras transmitidas mediante vibraciones. Se utilizaron señales vibrotáctiles generadas desde parlantes, y se estructuró el experimento en dos fases: entrenamiento y prueba. Los resultados revelaron que algunas palabras, como “amigo” y “mucho”, alcanzaron tasas de reconocimiento del 88%, mientras que otras, como “tarjeta”, “lluvia” y “calle”, apenas superaron el 55%. Esta disparidad podría explicarse por factores como la complejidad fonética y la similitud de las firmas vibrotáctiles, lo que sugiere que ciertas configuraciones acústicas son más fácilmente traducibles al canal táctil. Además, se evidenció una variabilidad significativa entre usuarios 69%–88%, posiblemente relacionada con diferencias individuales en sensibilidad, atención o estrategias de identificación.

El análisis comparativo entre ambos enfoques revela dos rutas complementarias hacia el reconocimiento de palabras: una, basada en el procesamiento profundo de señales acústicas; y otra, en la transducción sensorial mediante vibración. Si bien los modelos LSTM con características manuales demostraron una alta precisión en validación, el sistema vibrotáctil mostró ser funcional para transmitir información lingüística básica a través del tacto, lo cual abre posibilidades para aplicaciones inclusivas, especialmente en contextos de discapacidad auditiva o comunicación silenciosa. Sin embargo, la precisión táctil aún presenta limitaciones que deberán abordarse mediante técnicas de codificación más refinadas, entrenamiento prolongado de usuarios o mejoras en la resolución háptica del sistema.

5. Conclusiones Generales

El desarrollo de este proyecto se inició con una revisión de literatura científica, abarcando artículos, tesis y otros documentos académicos relacionados con la pérdida auditiva, los métodos de rehabilitación del habla, la clasificación de señales acústicas y la generación de señales vibrotáctiles. Esta revisión permitió identificar las técnicas más efectivas utilizadas en investigaciones previas como el uso de modelos preentrenados para la extracción de características de señales de audio, algoritmos de aprendizaje automático como redes neuronales recurrentes (RNN) y la aplicación de la envolvente multimodal para la generación de señales vibrotáctiles, lo que sirvió como base para la construcción del marco teórico y el diseño integral del sistema propuesto.

En el proyecto se diseñó una solución tecnológica integral para el reconocimiento de palabras habladas y su conversión en estímulos vibrotáctiles, orientada al apoyo en procesos de rehabilitación del habla. La arquitectura del sistema se construyó de forma modular, abarcando desde la adquisición, filtrado y normalización de señales de voz, hasta su clasificación mediante modelos HMM, LSTM y GRU. Se evaluaron en paralelo enfoques de extracción de características predefinidas (MFCC, deltas, espectrogramas y chroma) y métodos automáticos basados en el modelo preentrenado Wav2Vec2.0. La optimización de cada arquitectura se realizó utilizando el método de búsqueda de hiperparámetros de Optuna, garantizando un rendimiento superior al 85% y replicable en métricas como precisión, recall y F1-score.

En el enfoque de extracción de características predefinidas, las señales procesadas fueron segmentadas temporalmente y transformadas en vectores de características acústicas, normalizadas mediante Z-score. Estas características sirvieron como entrada para dos tipos de clasificadores: los Modelos Ocultos de Markov (HMM), que demostraron una capacidad aceptable de modelado secuencial, y redes neuronales LSTM, las cuales superaron ampliamente a los HMM en todas las métricas de evaluación (recall, precisión y F1-score). Específicamente, las arquitecturas LSTM optimizadas mediante Optuna mostraron un desempeño notable con F1-score de validación superiores al 0.95, gracias a su capacidad para modelar dependencias temporales complejas en las secuencias de voz.

Por otro lado, el enfoque basado en Wav2Vec2.0 permitió el uso directo de señales de voz crudas, las cuales fueron transformadas en vectores de alta dimensionalidad (768), encapsulando información fonética y prosódica de manera completa. Estos embeddings fueron empleados como entrada para modelos LSTM y GRU, también optimizados mediante búsqueda bayesiana con Optuna. Aunque estos modelos presentaron un rendimiento ligeramente inferior al de las LSTM con características manuales con un desempeño para F1-Score en validación de 0.86, ofrecieron una solución altamente escalable y menos dependiente de la ingeniería de características, lo cual representa una ventaja significativa en escenarios reales donde se requiere procesar grandes volúmenes de datos heterogéneos sin intervención manual.

En cuanto a la conversión a señales vibrotáctiles, se implementó un sistema de retroalimentación háptica capaz de transformar señales acústicas en estímulos

vibrotáctiles utilizando un banco de filtros Gammatone, transformada de Hilbert y modulación de amplitud con portadora de 150 Hz. Este sistema simuló la transducción coclear humana, permitiendo la percepción táctil de información fonética. Las pruebas experimentales con usuarios demostraron una precisión promedio del 74.8% ±7% en la identificación de palabras, destacándose "amigo" y "mucho" con tasas del 88% ±7% y ±6%. Esto valida la viabilidad del canal táctil como medio complementario de comunicación. mientras que otras como “tarjeta” mostraron mayores niveles de confusión, evidenciando la necesidad de optimización futura en el diseño de las señales hápticas.

Asimismo, se desarrolló una interfaz gráfica intuitiva y modular que permite al usuario grabar, visualizar, almacenar y comparar señales de voz, generando indicadores visuales y retroalimentación táctil, constituyéndose en una herramienta accesible para el entrenamiento y evaluación de pronunciación. Esta interfaz no solo facilita el uso del sistema en entornos de entrenamiento o rehabilitación, sino que también proporciona un entorno experimental controlado para la evaluación de algoritmos de procesamiento del habla.

En conjunto, los resultados de este proyecto demuestran la viabilidad técnica y funcional de un sistema híbrido de clasificación de palabras y retroalimentación táctil, que integra aprendizaje profundo, modelado acústico y diseño de interfaces. El enfoque permitió identificar que los modelos LSTM con características manuales alcanzan mejores métricas de clasificación en entornos controlados, mientras que los modelos basados en embeddings automáticos representan una alternativa más flexible y robusta para aplicaciones generalizables. A su vez, el sistema de estimulación vibrotáctil confirma que es posible transmitir contenido lingüístico a través del canal táctil, abriendo perspectivas para el desarrollo de tecnologías inclusivas orientadas a la accesibilidad comunicativa.

La limitante de número de grabaciones y número de usuarios seleccionadas para el entrenamiento de los modelos afectó su comprensión y clasificación para datos de nuevos hablantes. Además, la necesidad de ajustar los parámetros de los algoritmos en cada etapa del proceso fue una restricción computacional significativa, debido al elevado consumo de recursos y tiempo de procesamiento requerido. Este ajuste aseguró que cada parámetro sea optimizado para obtener mayor robustez en el sistema. Por otro lado, la generación de señales vibrotáctiles se vio afectada por el dispositivo de reproducción, el cual es capaz de reproducir señales de baja frecuencia, pero no con una gran intensidad, esto aumentó la dificultad en la percepción táctil durante las pruebas.

Finalmente, este trabajo demuestra que la integración de métodos computacionales avanzados con mecanismos sensoriales alternativos permite desarrollar soluciones efectivas, personalizables y accesibles para contextos terapéuticos, educativos y comunicación aumentativa. Esta propuesta abre nuevas líneas de investigación en el campo de las tecnologías hápticas aplicadas al lenguaje, consolidándose como una contribución significativa en la intersección entre el procesamiento de señales, inteligencia artificial y las tecnologías inclusivas.

6. Trabajos Futuros

Se proponen mejoras que podrían llevarse a cabo para avanzar en el desarrollo de sistemas que puedan ayudar en el proceso de rehabilitación del habla. Estas sugerencias están basadas en hallazgos y limitaciones identificadas durante el desarrollo de este proyecto.

Se recomienda la expansión y balanceo del conjunto de datos utilizado. Es fundamental contar con una base de datos más representativa, que incluya un mayor número de hablantes, distintas variantes dialectales del español, y una cobertura más amplia del vocabulario. Esto permitiría mejorar la generalización del sistema y facilitar su adaptación a distintos usuarios y contextos clínicos. Además, se sugiere incorporar grabaciones en condiciones acústicas variadas, simulando ambientes reales que incluyan ruido de fondo, reverberaciones u otros factores de interferencia, con el fin de robustecer el desempeño del sistema en escenarios no controlados.

El uso de RNNoise como técnica de mejora en la etapa de preprocesamiento de audio debido a su capacidad para reducir eficazmente el ruido de fondo sin comprometer la calidad del habla [59]. RNNoise es un sistema híbrido que combina procesamiento digital clásico con redes neuronales GRU, lo que le permite estimar de forma precisa las ganancias de atenuación en cada banda espectral. RNNoise se adapta dinámicamente a las características del ruido y puede funcionar en tiempo real con tasas de muestreo de hasta 48 kHz [60], lo que lo convierte en una solución eficiente para entornos ruidosos y sistemas embebidos.

Para la clasificación se plantea la exploración de arquitecturas más sofisticadas y métodos híbridos. Por ejemplo, la combinación de técnicas tradicionales como la Distorsión Dinámica del Tiempo (DTW) con modelos de aprendizaje profundo, particularmente redes neuronales profundas (DNN), podría optimizar el reconocimiento de comandos de voz en tiempo real. Este enfoque permite utilizar DTW para la alineación temporal precisa de señales de voz, mientras que la DNN se encargaría de la extracción y discriminación de características relevantes, mejorando la precisión del sistema en tareas de reconocimiento con vocabulario limitado [16].

Los descriptores utilizados para comparar la similitud entre audios pueden ampliarse mediante la incorporación no solo de las frecuencias de los formantes (F1 y F2), sino también de sus respectivas bandas de frecuencia (anchos de banda), con el fin de capturar de manera más completa las características acústico-fonéticas de la señal de voz. El uso combinado de formantes y bandas asociadas ha demostrado mejorar significativamente el desempeño en tareas de verificación de hablantes, al reflejar tanto las resonancias del tracto vocal como su estabilidad articulatoria y variabilidad espectral [54]. En este sentido, se propone que futuros desarrollos integren descriptores extendidos (F1, F2, BW1, BW2), lo cual permitiría evaluar la similitud entre señales de audio con mayor sensibilidad fonética.

Se puede explorar métodos que mejoren la generación de señales vibrotactiles, Se propone investigar la implementación de un vocoder táctil multibanda que incorpore

una etapa de expansión dinámica de la envolvente por bandas de frecuencia. Esta técnica busca amplificar selectivamente los formantes (asociados a vocales) y eventos transitorios (típicos de consonantes), incrementando así la percepción táctil diferenciada de los elementos fonéticos del habla. Dichas mejoras pueden tener un impacto directo en la discriminación táctil de palabras y frases, particularmente en usuarios con discapacidad auditiva [61]. Además de realizar pruebas con personas que cuenten con una discapacidad auditiva.

7. Referencias Bibliográficas

[1] “Hipoacusia | Sordera | PortalCLÍNIC.” Accessed: Dec. 19, 2024. [Online]. Available: https://www.clinicbarcelona.org/asistencia/enfermedades/sordera

[2] “Sordera y pérdida de la audición.” Accessed: Dec. 19, 2024. [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/deafness-and-hearing-loss

[3] “Pérdida de audición relacionada con la edad | NIDCD.” Accessed: Dec. 19, 2024. [Online]. Available: https://www.nidcd.nih.gov/es/espanol/perdida-de-audicion-relacionada-con-la-edad#ref6

[4] “Not Impossible Labs | Delve.” Accessed: Dec. 19, 2024. [Online]. Available: https://www.delve.com/work/not-impossible-labs

[5] Y. Quique B and M. FA, “Métodos unisensoriales para la rehabilitación de la persona con implante coclear y métodos musicoterapéuticos como nueva herramienta de intervención,” *Revista de otorrinolaringología y cirugía de cabeza y cuello*, vol. 73, no. 1, pp. 94–108, Apr. 2013, doi: 10.4067/S0718-48162013000100016.

[6] “(PDF) Interfaz Humano Máquina Audiovisual.” Accessed: Dec. 19, 2024. [Online]. Available: https://www.researchgate.net/publication/292986126_Interfaz_Humano_Maquina_Audiovisual#fullTextFileContent

[7] S. Camacho, D. Renza, and D. M. Ballesteros L, “A semi-supervised speaker identification method for audio forensics using cochleagrams,” *Communications in Computer and Information Science*, vol. 742, pp. 55–64, 2017, doi: 10.1007/978-3-319-66963-2_6/TABLES/2.

[8] L. Wanumen and H. Florez, “Architectural Approaches for Phonemes Recognition Systems,” *Communications in Computer and Information Science*, vol. 942, pp. 267–279, 2018, doi: 10.1007/978-3-030-01535-0_20/TABLES/3.

[9] J. O. Pinzón-Arenas, J. O. Pinzón-Arenas, and R. Jiménez-Moreno, “Comparison between handwritten word and speech record in real-time using CNN architectures,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, pp. 4313–4321, Aug. 2020, doi: 10.11591/ijece.v10i4.pp4313-4321.

[10] J. J. Herrera, R. Jimenez-Moreno, and J. E. M. Baquero, “Virtual environment for assistant mobile robot,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 6, pp. 6174–6184, Dec. 2023, doi: 10.11591/ijece.v13i6.pp6174-6184.

[11] D. Kumar and S. Aziz, “Performance Evaluation of Recurrent Neural Networks-

LSTM and GRU for Automatic Speech Recognition," *2023 International Conference on Computer, Electronics and Electrical Engineering and their Applications, IC2E3 2023*, 2023, doi: 10.1109/IC2E357697.2023.10262561.

[12] M. J. Luciá *et al.*, "Vibrotactile Captioning of Musical Effects in Audio-Visual Media as an Alternative for Deaf and Hard of Hearing People: An EEG Study," *IEEE Access*, vol. 8, pp. 190873–190881, 2020, doi: 10.1109/ACCESS.2020.3032229.

[13] R. V. Sharan, S. Berkovsky, and S. Liu, "Voice Command Recognition Using Biologically Inspired Time-Frequency Representation and Convolutional Neural Networks," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2020-July, pp. 998–1001, Jul. 2020, doi: 10.1109/EMBC44109.2020.9176006.

[14] I. Mykhailichenko, H. Ivashchenko, O. Barkovska, and O. Liashenko, "Application of Deep Neural Network for Real-Time Voice Command Recognition," *2022 IEEE 3rd KhPI Week on Advanced Technology, KhPI Week 2022 - Conference Proceedings*, 2022, doi: 10.1109/KHPIWEEK57572.2022.9916473.

[15] J. Cai, Y. Song, J. Wu, and X. Chen, "Voice Disorder Classification Using Wav2vec 2.0 Feature Extraction," *Journal of Voice*, Sep. 2024, doi: 10.1016/J.JVOICE.2024.09.002.

[16] Y. He, Y. Tao, and L. An, "Real-Time Voice Command Recognition System Based on DTW and MFCC," *Proceedings - 2023 6th International Conference on Computer Network, Electronic and Automation, ICCNEA 2023*, pp. 364–369, 2023, doi: 10.1109/ICCNEA60107.2023.00084.

[17] J. Oruh, S. Viriri, and A. Adegun, "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," *IEEE Access*, vol. 10, pp. 30069–30079, 2022, doi: 10.1109/ACCESS.2022.3159339.

[18] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, "Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review," *Electronics 2022, Vol. 11, Page 3795*, vol. 11, no. 22, p. 3795, Nov. 2022, doi: 10.3390/ELECTRONICS11223795.

[19] Y. Shao, V. Hayward, and Y. Visell, "Spatial patterns of cutaneous vibration during whole-hand haptic interactions," *Proc Natl Acad Sci U S A*, vol. 113, no. 15, pp. 4188–4193, Apr. 2016, doi: 10.1073/PNAS.1520866113/SUPPL_FILE/PNAS.201520866SI.PDF.

[20] I. S. Răutu, X. De Tiège, V. Jousmäki, M. Bourguignon, and J. Bertels, "Speech-derived haptic stimulation enhances speech recognition in a multi-talker background," *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–11, Oct. 2023, doi: 10.1038/S41598-023-43644-3.

[21] M. Sobhan, M. Z. Chowdhury, I. Ahsan, H. Mahmu, and M. K. Hasan, “A Communication Aid System for Deaf and Mute using Vibrotactile and Visual Feedback,” *Proceedings - 2019 International Seminar on Application for Technology of Information and Communication: Industry 4.0: Retrospect, Prospect, and Challenges, iSemantic 2019*, pp. 184–190, Sep. 2019, doi: 10.1109/ISEMANTIC.2019.8884323.

[22] Balakrishnan.K, K. .P, S. .K, and S. . M. .V, “Embedded Based Vibration Motor Using Deaf and Dumb Peoples,” *International Journal of Engineering Research & Technology*, vol. 12, no. 3, May 2024, doi: 10.17577/IJERTCONV12IS03094.

[23] Á. García López, V. Cerdán, T. Ortiz, J. M. Sánchez Pena, and R. Vergaz, “Emotion Elicitation through Vibrotactile Stimulation as an Alternative for Deaf and Hard of Hearing People: An EEG Study,” *Electronics 2022, Vol. 11, Page 2196*, vol. 11, no. 14, p. 2196, Jul. 2022, doi: 10.3390/ELECTRONICS11142196.

[24] L. Turchet, R. Rosaia, A. Diodati, and M. Carner, “Exposure to vibrotactile music improves audiometric performances in individuals with cochlear implants,” *Sci Rep*, vol. 15, no. 1, pp. 1–11, Dec. 2025, doi: 10.1038/S41598-025-02946-4;SUBJMETA.

[25] A. Schulte, J. Marozeau, A. Ruhe, A. Büchner, A. Kral, and H. Innes-Brown, “Improved speech intelligibility in the presence of congruent vibrotactile speech input,” *Sci Rep*, vol. 13, no. 1, pp. 1–17, Dec. 2023, doi: 10.1038/S41598-023-48893-W;SUBJMETA.

[26] M. Mirzaei, P. Kán, and H. Kaufmann, “Effects of using vibrotactile feedback on sound localization by deaf and hard-of-hearing people in virtual environments,” *Electronics (Switzerland)*, vol. 10, no. 22, Nov. 2021, doi: 10.3390/ELECTRONICS10222794.

[27] “Types of Hearing Loss | Hearing Loss in Children | CDC.” Accessed: Dec. 19, 2024. [Online]. Available: https://www.cdc.gov/hearing-loss-children/about/types-of-hearing-loss.html?CDC_AAref_Val=https://www.cdc.gov/ncbddd/hearingloss/types.html

[28] M. Maggiolo and F. E. Schwalm, “Test de articulación a la repetición (TAR): un legado de la profesora fonoaudióloga Edith Schwalm,” *Revista Chilena de Fonoaudiología*, vol. 16, pp. 1–14, Nov. 2017, doi: 10.5354/0719-4692.2017.47557.

[29] S. A. Abad, N. Herzig, D. Raitt, M. Koltzenburg, and H. Wurdemann, “Bioinspired adaptable multiplanar mechano-vibrotactile haptic system,” *Nature Communications 2024 15:1*, vol. 15, no. 1, pp. 1–12, Sep. 2024, doi: 10.1038/S41467-024-51779-8.

[30] L. R. Rabiner and R. W. Schafer, “the essence of knowledge Introduction to

Digital Speech Processing Introduction to Digital Speech Processing".

[31] S. Escobar-Pajoy and J. P. Ugarte, "Computerized analysis of pulmonary sounds using uniform manifold projection," *Chaos Solitons Fractals*, vol. 166, p. 112930, Jan. 2023, doi: 10.1016/J.CHAOS.2022.112930.

[32] Y. Huang, B. Li, M. Barni, and J. Huang, "Identification of VoIP Speech with Multiple Domain Deep Features," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2253–2267, 2020, doi: 10.1109/TIFS.2019.2960635.

[33] R. Sun, "A Design Method of Active High-Pass Butter-worth Filter," *IEEE International Workshop on Electromagnetics: Applications and Student Innovation Competition, iWEM 2021 - Proceedings*, 2021, doi: 10.1109/IWEM53379.2021.9790408.

[34] U. Profesional, A. López Mateos, E. Yidell Carbajal Degante Asesores, and M. C. en Eric Gómez Gómez Ing Carlos Mira González México, "INSTITUTO POLITÉCNICO NACIONAL ESCUELA SUPERIOR DE INGENIERÍA MECÁNICA Y ELÉCTRICA "REDUCCIÓN DE RUIDO EN SEÑ".

[35] A. V. . Oppenheim, R. W. . Schafer, and J. R. . Buck, "Discrete-time signal processing," p. 864, 20061999, Accessed: Dec. 19, 2024. [Online]. Available: https://books.google.com/books/about/Discrete_Time_Signal_Processing.html?hl=es&id=geTn5W47KEsC

[36] S. Hizlisoy, R. S. Arslan, and E. Çolakoğlu, "Singer identification model using data augmentation and enhanced feature conversion with hybrid feature vector and machine learning," *EURASIP J Audio Speech Music Process*, vol. 2024, no. 1, pp. 1–13, Dec. 2024, doi: 10.1186/S13636-024-00336-8/TABLES/3.

[37] "¿Qué es el aprendizaje automático? Una definición — Aprendizaje automático — DATA SCIENCE." Accessed: Dec. 19, 2024. [Online]. Available: https://datascience.eu/es/aprendizaje-automatico/que-es-el-aprendizaje-automatico-una-definicion/

[38] D. G. . Stork, "Pattern Classification (2nd ed.)," Jan. 01, 1999. Accessed: Dec. 19, 2024. [Online]. Available: https://www.academia.edu/110850149/Pattern_Classification_2nd_ed_

[39] "¿Qué es el aprendizaje supervisado? | IBM." Accessed: Dec. 19, 2024. [Online]. Available: https://www.ibm.com/es-es/topics/supervised-learning

[40] J. Díaz-Ramírez and J. Díaz-Ramírez, "Machine Learning and Deep Learning," *Ingeniare. Revista chilena de ingeniería*, vol. 29, no. 2, pp. 180–181, Jun. 2021, doi: 10.4067/S0718-33052021000200180.

[41] "Speech and Language Processing." Accessed: Dec. 19, 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[42] C. M. Bishop, *Prml*. 2006. Accessed: Dec. 19, 2024. [Online]. Available: https://link.springer.com/book/9780387310732

[43] B. Kommey, E. O. Addo, and E. Tamakloe, “A Hidden Markov Model-Based Speech Recognition System Using Baum-Welch, Forward-Backward and Viterbi Algorithms,” *Jordan Journal of Electrical Engineering*, vol. 9, no. 4, pp. 509–536, 2023, doi: 10.5455/JJEE.204-1675950756.

[44] A. Kuamr, M. Dua, and T. Choudhary, “Continuous hindi speech recognition using Gaussian mixture HMM,” *2014 IEEE Students’ Conference on Electrical, Electronics and Computer Science, SCEECS 2014*, 2014, doi: 10.1109/SCEECS.2014.6804519.

[45] T. M. Maina, “Optimizing Phonetic Recognition and Computational Efficiency in Swahili Digraphs Using Feature Reduction Model with Multinomial Logistic Regression,” *World Academics Journal of Engineering Sciences*, vol. 12, no. 1, pp. 16–25, 2025, Accessed: May 27, 2025. [Online]. Available: www.isroset.org

[46] G. A. Martínez Mascorro and G. Aguilar Torres, “Reconocimiento de voz basado en MFCC, SBC y Espectrogramas,” *Ingenius*, no. 10, Dec. 2013, doi: 10.17163/INGS.N10.2013.02.

[47] “(PDF) Chroma Feature Extraction.” Accessed: Dec. 24, 2024. [Online]. Available: https://www.researchgate.net/publication/330796993_Chroma_Feature_Extraction

[48] D. W. Otter, J. R. Medina, and J. K. Kalita, “A Survey of the Usages of Deep Learning for Natural Language Processing,” *IEEE Trans Neural Netw Learn Syst*, vol. 32, no. 2, pp. 604–624, Feb. 2021, doi: 10.1109/TNNLS.2020.2979670.

[49] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, “Improving speech recognition by revising gated recurrent units,” Sep. 2017, Accessed: Dec. 19, 2024. [Online]. Available: http://arxiv.org/abs/1710.00641

[50] H. Kuttruff, “Room acoustics, sixth edition,” *Room Acoustics, Sixth Edition*, pp. 1–302, Oct. 2016, doi: 10.1201/9781315372150/ROOM-ACOUSTICS-HEINRICH-KUTTRUFF/ACCESSIBILITY-INFORMATION.

[51] R. Islam and M. Tarique, “Investigating the Performance of Gammatone Filters and Their Applicability to Design Cochlear Implant Processing System,” *Designs 2024, Vol. 8, Page 16*, vol. 8, no. 1, p. 16, Feb. 2024, doi: 10.3390/DESIGNS8010016.

[52] Q. Ding, X. L. Jiang, H. Q. Huang, J. J. Cai, and W. Su, “Estimation of code-width based on Hilbert transform,” *2017 IEEE International Conference on Signal Processing, Communications and Computing, ICSPCC 2017*, vol. 2017-January, pp. 1–4, Dec. 2017, doi: 10.1109/ICSPCC.2017.8242454.

[53] D. Gowda, S. R. Kadiri, B. Story, and P. Alku, “Time-Varying Quasi-Closed-Phase Analysis for Accurate Formant Tracking in Speech Signals,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 28, pp. 1901–1914, 2020, doi: 10.1109/TASLP.2020.3000037.

[54] T. Becker, M. Jessen, and C. Grigoras, “Forensic speaker verification using formant features and Gaussian mixture models,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1505–1508, 2008, doi: 10.21437/INTERSPEECH.2008-432.

[55] R. Wainschenker, J. H. Doorn, M. Castro, and C. F. Legrottaglie, “Cálculo y análisis del pitch en señales sonoras de voz humana,” 2003, Accessed: Dec. 19, 2024. [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/21513

[56] “(PDF) Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation.” Accessed: Dec. 19, 2024. [Online]. Available: https://www.researchgate.net/publication/276412348_Evaluation_From_precision_recall_and_F-measure_to_ROC_informedness_markedness_correlation

[57] H. Loaiza, “Matriz de Confusión y Métricas,” 2023.

[58] “openslr.org.” Accessed: May 25, 2025. [Online]. Available: https://www.openslr.org/72/

[59] A. S. Shkil, O. I. Filippenko, D. Y. Rakhlis, I. V. Filippenko, A. V. Parkhomenko, and V. R. Korniienko, “ADAPTIVE FILTERING AND MACHINE LEARNING METHODS IN NOISE SUPPRESSION SYSTEMS, IMPLEMENTED ON THE SoC,” *Radio Electronics, Computer Science, Control*, no. 4, pp. 163–174, Dec. 2024, doi: 10.15588/1607-3274-2024-4-16.

[60] A. Kanev, M. Nazarov, D. Uskov, and V. Terentyev, “Research of Different Neural Network Architectures for Audio and Video Denoising,” *Proceedings of the 2023 5th International Youth Conference on Radio Electronics, Electrical and Power Engineering, REEPE 2023*, 2023, doi: 10.1109/REEPE57272.2023.10086862.

[61] M. D. Fletcher, E. Akis, C. A. Verschuur, and S. W. Perry, “Improved tactile speech perception and noise robustness using audio-to-tactile sensory substitution with amplitude envelope expansion,” *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1–12, Jul. 2024, doi: 10.1038/S41598-024-65510-6.

8. Anexos

Anexo 1. Codigo de la interfaz

Para acceder al código de la interfaz se debe visitar el repositorio de GitHub: https://github.com/Ajota3744/Dise-oInterfaz.git

Anexo 2. Manual de usuario interfaz

A2.1. Descripción general

La interfaz gráfica desarrollada permite registrar usuarios, grabar señales de voz organizadas por palabras y fonemas, visualizar y reproducir grabaciones, generar señales vibrotáctiles, y comparar la similitud acústica con grabaciones de referencia. Está diseñada para ser intuitiva, adaptable y adecuada para experimentos de reconocimiento de voz o estimulación sensorial.

A2.2. Requisitos de uso

- Sistema operativo: Windows, Linux o macOS
- Micrófono conectado y funcional
- Python 3.8 o superior con bibliotecas instaladas: PyQt5, librosa, matplotlib, scipy, numpy, sounddevice, parselmouth, noisereduce.

A2.3. Instrucciones de uso

A2.3.1 Registro de usuario

1. Escriba su nombre, apellido y edad.
2. Seleccione el género (Masculino o Femenino).
3. Presione "Registrar usuario". Se creará automáticamente una carpeta bajo Usuarios/Nombre Apellido Edad Género.

A2.3.2 Selección de palabra

1. Vaya al menú superior Fonemas → Palabra.
2. Al seleccionar una palabra:

- Se mostrará en bajo la señal de referencia
- Se cargará su imagen correspondiente en señal de referencia
- Se graficará el audio de referencia correspondiente al género del usuario

A2.3.3 Grabación de audio

1. Presione el botón "Grabar" para iniciar la captura.
2. Vuelva a presionar para detener.
3. El archivo será preprocesado (filtro + reducción de ruido) y guardado en la carpeta del usuario según el fonema y palabra seleccionados.
4. La señal grabada se grafica automáticamente en señal de usuario.

A2.3.4. Reproducción de grabaciones

Puede reproducir el último audio grabado o una grabación seleccionada desde el treeView con el botón "Reproducir".

A2.3.5 Visualización y gestión de grabaciones

1. Presione "Cargar usuario" y seleccione su carpeta.
2. Se cargarán las grabaciones en el panel de usuario.
3. Haga doble clic sobre un archivo .wav para graficarlo.
4. Puede hacer clic derecho sobre una grabación para eliminarla permanentemente.

A2.3.6. Generación de señal vibrotáctil

1. Seleccione una grabación (grabada o desde el panel).
2. Presione “Vibraciones”.
3. Se calculará la envolvente multibanda y se generará una señal modulada en amplitud (150 Hz), que se reproduce y grafica.

A2.3.7 Comparación con referencia

1. Seleccione una palabra desde el menú Fonemas.
2. Asegúrese de tener una grabación reciente o seleccionada.
3. Presione "Comparar con referencia".
4. El sistema extrae características (energía, pitch, formantes) de ambas señales, calcula su similitud y:

- Muestra el resultado en la barra de progreso con color (rojo/amarillo/verde).

- Muestra un mensaje textual con interpretación del resultado (baja, aceptable, buena similitud).

A2.4. Precauciones

1. No elimine manualmente archivos dentro de la carpeta Usuarios, hágalo desde la interfaz.

2. Si el micrófono no está conectado, la grabación puede fallar.

3. La comparación acústica requiere que exista un archivo de referencia en la carpeta correspondiente de Referencias.

Anexo 3. Consentimiento informado

Para garantizar la participación ética y voluntaria en el desarrollo del proyecto, se elaboró un formato de consentimiento informado que describe los objetivos, alcances y posibles implicaciones de la investigación. Dicho documento fue presentado a las seis personas participantes, quienes lo leyeron y aceptaron de manera libre e informada, manifestando por escrito su conformidad para vincularse en las actividades propuestas.

Universidad del Valle

CONSENTIMIENTO INFORMADO

PARTICIPANTE

SISTEMA AUDIOVISUAL DE AYUDAS PARA PERSONAS SRODAS

Código del participante: Participante 1

Lo estamos invitando a participar en una investigación que tiene como objetivo evaluar el desempeño de una interfaz en la cual participarán 8 personas incluido(a) usted. Le garantizamos que ni su nombre ni sus datos personales serán expuestos, ya que de aquí en adelante solo se identificará con el código a usted asignado en la parte superior de este documento.

Su participación en el estudio es totalmente voluntaria y consiste en: La prueba del sistema de reconocimiento de voz de la interfaz para 12 palabras ya seleccionadas. Su participación tiene un tiempo aproximado de duración de 20 minutos. Es importante aclarar que en caso no darse completamente las indicaciones aquí explicadas, se le retirará del estudio sin tener ninguna repercusión en la en la atención recibida ni en la actividad que realice el participante. Otra forma de retirarse del proyecto es de manera voluntaria, para lo cual solo debe expresar su deseo sin que esto ocasione ningún problema para usted. Otra forma de retirarse del proyecto es de manera voluntaria, para lo cual solo debe expresar su deseo sin que esto ocasione ningún problema para usted.

Su participación puede ocasionarle algunos riesgos como son fatiga o incomodidad por el uso prolongado de la voz, los cuales serán evitados mediante pausas entre repeticiones, duración limitada de la sesión y monitoreo constante del bienestar del participante. En caso de presentarse algún inconveniente contamos con el personal de la salud del centro universitario adecuado para solucionarlo.

Los beneficios que obtendrá serán la experiencia directa con una tecnología de reconocimiento de voz, y la posibilidad de contribuir al desarrollo de sistemas accesibles y eficientes para la comunidad, y debe saber que no incurrirá en ningún gasto por cuenta de su participación y tampoco recibirá ninguna compensación económica por la misma.

La información obtenida en este estudio será guardada mediante un código que permitirá que nadie ajeno a la investigación acceda a su información privada.

Le informamos que una vez termine el estudio le estaremos dando a conocer los resultados.

En caso de solicitar información adicional puede comunicarse con el Dr. Humberto Loaiza al Tel.: +57 2 3391780 Ext. 7654 o con Anthony Muñoz al Cel 3126192471. Si requiere información sobre los aspectos éticos de su participación puede comunicarse con el Comité de Ética en Investigación en Salud – CEIS - al teléfono (602) 5185677 o al correo eticasalud@correounivalle.edu.co.

Igualmente le solicitamos que usted de su Autorización para que las grabaciones de voz sean usadas en los siguientes procesos, guardando su identidad y con la previa autorización de un comité de ética. Marque con una X según decida:

- Autorizo el uso en actividades con fines académicos SI __x__ NO____
- Autorizo el uso en futuros estudios SI _x__ NO____

Se entrega una copia de este documento al participante.

Nombre del participante: Anthony Muñoz
Firma: Anthony Muñoz

Persona que realizó el proceso para obtener el consentimiento informado
Nombre: Anthony Muñoz Correa
CC: 1061806725
Firma: Anthony Muñoz
Relación con la investigación: Autor

Se firma en Cali, a los 29 días del mes de mayo del 2025