



Universidad Autónoma del Estado de Hidalgo
Instituto de Ciencias Básicas e Ingeniería
Licenciatura en Ciencias Computacionales

**Reconocimiento de hablantes mediante
representaciones unidimensionales y
bidimensionales de señales de voz usando
aprendizaje automático**

T E S I S

Que para obtener el título de
Licenciado en Ciencias Computacionales

Presenta

José Isac Frausto Cadena

Directores de Tesis

Dra. Rosa María Ortega Mendoza

Dr. Ricardo Calderón Suárez

Mineral de la Reforma, Hidalgo.
Julio de 2026



Universidad Autónoma del Estado de Hidalgo

Instituto de Ciencias Básicas e Ingeniería

School of Engineering and Basic Sciences

Dirección

Office of the Director

Mineral de la Reforma, Hgo., a 30 de julio de 2026

Número de control: ICBI-D/1333/2026

Asunto: Autorización de impresión.

MTRA. OJUKY DEL ROCÍO ISLAS MALDONADO
DIRECTORA DE ADMINISTRACIÓN ESCOLAR DE LA UA EH

Con Título Quinto, Capítulo II, Capítulo V, Artículo 51 Fracción IX del Estatuto General de nuestra Institución, por este medio, le comunico que el Jurado asignado al egresado de la Licenciatura en Ciencias Computacionales **José Isac Frausto Cadena**, quien presenta el trabajo de titulación **"Reconocimiento de hablantes mediante representaciones unidimensionales y bidimensionales de señales de voz usando aprendizaje automático"**, ha decidido, después de revisar fundamento en lo dispuesto en el Título Tercero, Capítulo I, Artículo 18 Fracción IV; dicho trabajo en la reunión de sinodales, **autorizar la impresión del mismo**, una vez realizadas las correcciones acordadas.

A continuación, firman de conformidad los integrantes del Jurado:

Presidente: Dr. Félix Agustín Castro Espinoza

Secretario: Dra. Guadalupe Carmona Arroyo

Vocal: Dra. Rosa María Ortega Mendoza

Suplente: Dr. Ricardo Calderón Suárez

Sin otro particular por el momento, reciba un cordial saludo.

Atentamente
"Amor, Orden y Progreso"

Dr. Gabriel Vergara Rodríguez
Director del ICBI



GVR/MMM

"Amor, Orden y Progreso"

Ciudad del Conocimiento, Carretera Pachuca-Tulancingo
Km. 4.5 Colonia Carboneras, Mineral de la Reforma, Hidalgo,
México. C.P 42184

Teléfono: 52 (771) 71 720 00 Ext. 40001

direccion_icbi@uaeh.edu.mx, vergara@uaeh.edu.mx



uaeh.edu.mx

Dedicatoria

A mis papás, Norma y José Luis, por darme tantas oportunidades, muchas más de las que ustedes tuvieron en su momento. Gracias por permitirme estudiar, por brindarme su ayuda incondicional cuando más la necesité y por estar presentes en cada etapa importante. Este logro representa directamente su sacrificio y es el resultado de años de su trabajo. Gracias por dar todo de sí para que yo pudiera llegar hasta aquí, por ser mi casa, mi apoyo y mi fuerza. A mi hermano Alan, por ser un guía a lo largo de mi vida y por mostrarme, a través del ejemplo de sus propios aciertos y aprendizajes, el camino para tomar mis propias decisiones. *¡Gracias familia!*

A mis amigos, Edwin Campos Dragusin y Cesar Castelazo Ortiz, con quienes cursé toda la carrera y a quienes les agradezco haberme acompañado. Fuimos el mejor equipo dentro de los desafíos que significaron las entregas de último momento y los proyectos que parecían imposibles, pero también compartimos muchísimos momentos divertidos que marcaron mi vida para bien. Espero que mantengamos por mucho tiempo nuestra buena amistad.

A mi pareja, Dafne, agradezco mucho haberte conocido. Gracias por acompañarme en este proceso, por darme alegría y motivación cuando más lo requería, por apoyarme de tantas maneras y por hacer el esfuerzo de estar conmigo incluso en los días que más cansada te sentías. Tu presencia fue muy importante para mí a lo largo de este camino. Este trabajo también es en parte tuyo, muchas gracias por hacer de mi vida algo mucho mejor y por darme la fortuna de compartir tu cariño y amor.

Y finalmente, a mi querida Fufi, que ahora descansa en paz. Gracias desde lo más profundo de mi ser por acompañarme en incontables noches de desvelo, por tu presencia silenciosa y tu amor incondicional. Tu ausencia me duele cada día y te extraño mucho, te llevaré siempre en mi corazón. Te mando un fuerte abrazo y muchas caricias hasta donde estés. Te agradezco infinitamente por haber sido parte de mi vida.

Agradecimientos

Deseo expresar mi sincero agradecimiento a quienes hicieron posible este trabajo desde el ámbito académico.

A la Dra. Rosa María Ortega Mendoza, por su guía, apoyo, compromiso y acompañamiento constante durante el desarrollo de esta tesis. Su orientación fue fundamental para consolidar este proyecto.

Al Dr. Ricardo Calderón Suárez, por sus valiosas observaciones, sugerencias metodológicas y disposición para enriquecer este trabajo con su experiencia.

A los sinodales que evaluaron esta tesis, por su tiempo, disposición y aportaciones que sin duda contribuyeron a fortalecer la calidad del presente documento.

A todos aquellos profesores y profesoras que formaron parte de mi trayectoria universitaria y que realmente buscaron impartir su conocimiento, gracias por fomentar en mí el pensamiento crítico y la pasión por aprender.

A la Universidad Autónoma del Estado de Hidalgo, por brindarme las herramientas, el entorno académico y la formación integral que hicieron posible la culminación de esta etapa.

Resumen

El presente trabajo tiene como objetivo evaluar y comparar distintas representaciones de la señal de voz para la tarea de reconocimiento de hablantes, considerando tanto un enfoque tradicional unidimensional basado en los coeficientes cepstrales en la escala de Mel (MFCC), como una representación bidimensional derivada de los mismos coeficientes en forma de imágenes.

En el estudio se realizaron diversos experimentos orientados a analizar los factores que influyen en el desempeño del modelo y en la estabilidad de los resultados. Se evaluaron diferentes configuraciones que incluyeron el impacto de las técnicas de normalización y estandarización, la variación del número de coeficientes MFCC, la incorporación de coeficientes delta, la reducción dimensional del espacio de atributos mediante ganancia de información y la influencia del contenido verbal.

En la etapa final se estableció una comparación entre representaciones unidimensionales y bidimensionales bajo un marco experimental común, considerando la misma tarea de clasificación, características base y métricas de evaluación. Se evaluó el desempeño de clasificadores tradicionales basados en vectores, como máquinas de vectores de soporte, k-vecinos más cercanos y bosques aleatorios, contrastándose frente a una arquitectura profunda VGG16 entrenada sobre las representaciones bidimensionales de los mismos coeficientes acústicos transformados en imágenes. Los resultados mostraron la estabilidad y consistencia que mantiene el enfoque unidimensional en conjuntos de datos más reducidos, mientras que la representación bidimensional permitió explorar la sensibilidad de la red convolucional a los datos, presentando un incremento considerable en su capacidad discriminativa conforme se dispone de mayor cantidad de información. De esta forma, la experimentación evidencia tanto la eficiencia del enfoque vectorial como el potencial del enfoque visual en escenarios de mayor complejidad multiclase.

En conclusión, la comparación entre representaciones unidimensionales y bidimensionales indica que los MFCC en formato vectorial de dimensión fija continúan siendo una alternativa sólida y eficiente para las tareas de reconocimiento de hablantes, especialmente en contextos con conjuntos de datos moderados. No obstante, las representaciones bidimensionales ofrecen un ca-

mino prometedor para capturar estructuras espaciales complejas de la señal de voz, siempre que se disponga de un volumen de datos suficiente para el aprendizaje de las capas convolucionales y una adecuada optimización de hiperparámetros. Esta diferencia indica la importancia de seleccionar la representación en función del escenario experimental y abre la posibilidad de integrar ambos enfoques en futuros desarrollos multimodales.

Palabras clave: Reconocimiento de hablantes, coeficientes cepstrales en la escala de Mel, máquinas de vectores de soporte, redes neuronales convolucionales, VGG16, representación de características.

Abstract

The objective of this work is to evaluate and compare different speech signal representations for the speaker recognition task, considering both a traditional one-dimensional approach based on Mel-Frequency Cepstral Coefficients (MFCC) and a two-dimensional representation derived from the same coefficients in the form of images.

Several experiments were conducted in this study to analyze the factors influencing model performance and result stability. Different configurations were evaluated, including the impact of normalization and standardization techniques, variations in the number of MFCCs, the incorporation of delta coefficients, dimensional reduction of the feature space through information gain, and the influence of verbal content.

In the final stage, a direct comparison between one-dimensional and two-dimensional representations was established under the same experimental framework. The performance of traditional vector-based classifiers, such as support vector machines, k-nearest neighbors, and random forests, was evaluated and contrasted against a VGG16 deep architecture trained on the two-dimensional representations of the same acoustic coefficients transformed into images. The results showed the stability and consistency maintained by the one-dimensional approach on smaller datasets, whereas the two-dimensional representation allowed exploring the sensitivity of the convolutional network to data, demonstrating a considerable increase in its discriminative capacity as more information becomes available. Thus, the experimentation evidences both the efficiency of the vector approach and the potential of the visual approach in scenarios of higher multi-class complexity.

In conclusion, the comparison between one-dimensional and two-dimensional representations demonstrates that MFCCs in a fixed-dimension vector format remain a solid and efficient alternative for speaker recognition tasks, particularly in contexts with moderate datasets. Nonetheless, two-dimensional representations offer a promising path to capture complex spatial structures of the speech signal, provided that a sufficient volume of data is available for the learning of convolutional layers and an appropriate hyperparameter optimization. This difference indicates the

importance of selecting the representation based on the experimental scenario and opens the possibility of integrating both approaches in future multimodal developments.

Keywords: Speaker recognition, Mel-Frequency Cepstral Coefficients, Support Vector Machines, Convolutional Neural Networks, VGG16, Feature representation.

Índice general

Dedicatoria	II
Agradecimientos	IV
Resumen	VI
Abstract	VIII
1. Introducción	1
1.1. Planteamiento del problema	2
1.2. Justificación	3
1.2.1. Objetivo general	4
1.2.2. Objetivos específicos	5
1.3. Solución propuesta	5
1.4. Organización de la tesis	6
2. Marco conceptual	9
2.1. La señal de voz	9
2.2. Características acústicas de la voz	10
2.2.1. Tono	10
2.2.2. Timbre	10
2.2.3. Intensidad	11
2.2.4. Duración	11
2.2.5. Formantes	11
2.3. Señal analógica y digital	11
2.3.1. Importancia en el procesamiento de voz	14
2.4. Procesamiento digital de señales de voz	15
2.4.1. Dificultades y desafíos en el procesamiento digital de señales de voz	17

2.4.2.	Frecuencia y percepción auditiva	18
2.4.3.	Escala de Mel	22
2.5.	Coeficientes cepstrales en la escala de Mel (MFCC)	25
2.5.1.	Delta y delta-delta	29
2.6.	Procesamiento del lenguaje natural (PLN) y su relación con el reconocimiento de hablantes	30
2.6.1.	Reconocimiento del hablante	31
2.6.2.	Diferencia entre reconocimiento de hablante y de voz	33
2.6.3.	Representaciones de coeficientes MFCC usando espacios 1D y 2D	34
2.7.	Aprendizaje automático y clasificación	36
2.7.1.	Aprendizaje supervisado	37
2.7.2.	Algoritmos de aprendizaje automático	39
2.8.	Redes neuronales y aprendizaje profundo	46
2.8.1.	Redes neuronales artificiales	46
2.8.2.	Aprendizaje profundo	48
3.	Estado del arte	55
3.1.	Enfoques de clasificación supervisada para reconocer al hablante	55
3.1.1.	Enfoques basados en representaciones en formato 1D	55
3.1.2.	Enfoques basados en representaciones en formato 2D	61
3.1.3.	Comparación de representaciones acústicas multimodales	68
3.2.	Conjuntos de datos	70
3.2.1.	Conjunto de datos DIMEX100	70
3.2.2.	Conjunto de datos VoxCeleb1	70
3.2.3.	Estadísticas de los conjuntos de datos	71
3.3.	Discusión	71
4.	Metodología	77

4.1.	Etapa 1: Extracción de características	78
4.1.1.	Búsqueda y selección de los conjuntos de datos	78
4.1.2.	Preprocesamiento avanzado de la señal de audio	78
4.1.3.	Extracción de MFCC	79
4.2.	Etapa 2: Clasificación	80
4.2.1.	Construcción de representaciones	83
4.2.2.	Construcción de modelos computacionales de clasificación	84
4.2.3.	Estrategia del bloque experimental	85
5.	Experimentos y resultados	87
5.1.	Configuración general	87
5.2.	Experimentos usando representaciones 1D	89
5.2.1.	Experimento 1. Comparación del impacto de la normalización y estandarización en las características MFCC	89
5.2.2.	Experimento 2. Evaluación de diferentes algoritmos de clasificación	90
5.2.3.	Experimento 3. Evaluación del uso de deltas como características de la representación	92
5.2.4.	Experimento 4. Uso de selección de características	92
5.2.5.	Experimento 5. Audios comunes vs individuales (DIMEX)	94
5.3.	Experimentos usando representaciones 2D	95
5.3.1.	Experimento 6. Representación bidimensional (imágenes de MFCC) con tres clases	95
5.3.2.	Experimento 7. Representación bidimensional con cinco clases	96
5.3.3.	Experimento 8. Representación bidimensional con cinco clases y 40 coeficientes MFCC	97
5.4.	Comparación de 1D y 2D	98
5.5.	Análisis de resultados y hallazgos clave	102

6. Conclusiones y trabajo futuro	105
6.1. Conclusiones	105
6.2. Trabajo futuro	107

Índice de figuras

2.1. Señal analógica	12
2.2. Señal digital	13
2.3. Representación de una señal discreta	14
2.4. Etapas del preprocesamiento de señales de voz para reconocimiento del hablante .	17
2.5. Representación de una frecuencia alta y baja.	19
2.6. Curvas de respuesta en frecuencia, umbrales de audición y dolor del oído humano .	20
2.7. Representación de la escala musical en el sistema anglosajón	21
2.8. Resultados del experimento de fraccionamiento tonal de Stevens	23
2.9. Gráfica de la escala Mel	24
2.10. Segmentación en marcos de la señal de voz	26
2.11. Función de ventana de Hamming	27
2.12. Representación de la señal en el dominio del tiempo y la frecuencia	28
2.13. Verificación del hablante	32
2.14. Representación 2D	36
2.15. Aprendizaje supervisado	38
2.16. Hiperplano	40
2.17. Funcionamiento de los bosques aleatorios	43
2.18. Funcionamiento del algoritmo k-vecinos más cercanos	45
2.19. Red neuronal artificial	48
2.20. Estructura típica de una arquitectura CNN	51
2.21. Estructura interna de la arquitectura VGG16	53
3.1. Flujo general del modelo 1D-Triplet-CNN de Chowdhury y Ross (2020)	57
3.2. Rendimiento con método tradicional y parámetros aprendibles de la propuesta de Liu et al. (2021)	59
3.3. Sistema de reconocimiento del hablante adaptado de Zhang y Liu (2023)	60

3.4.	Sistema de reconocimiento del hablante adaptado de Wirdiani et al. (2024)	63
3.5.	Sistema de reconocimiento del hablante adaptado de Youngsik y Bang (2021) . . .	65
3.6.	Arquitectura de la red neuronal convolucional CNN1 adaptada de Costantini et al. (2023)	67
3.7.	Representaciones visuales en formato de imagen física obtenidas de Costantini et al. (2023)	67
4.1.	Metodología general	77
4.2.	Diagrama de flujo de la etapa dos de la metodología para señales de audio 1D . . .	81
4.3.	Diagrama de flujo de la etapa dos de la metodología para señales de audio 2D . . .	82
5.1.	Comparación de la exactitud entre datos originales y reducidos	93

Índice de Tablas

3.1. Medidas estadísticas aplicadas a los cambios de los coeficientes. Adaptado (traducido) de Ortega-Mendoza et al. (2007).	56
3.2. Resumen de los conjuntos de datos utilizados	71
3.3. Comparativa de trabajos en el estado del arte sobre representaciones acústicas para el reconocimiento de hablantes	74
5.1. Métricas promedio por conjunto de datos y método de preprocesamiento	90
5.2. Resultados de la evaluación de clasificadores bajo la variación en la cantidad de coeficientes MFCC	91
5.3. Métricas promedio por representación	92
5.4. Resultados de reducción de atributos en el conjunto de datos DIMEX	93
5.5. Rendimiento del modelo SVM con audios comunes e individuales	94
5.6. Métricas promedio para la representación bidimensional con tres clases y 100 coeficientes MFCC	95
5.7. Métricas promedio para la representación bidimensional con cinco clases y 100 coeficientes MFCC	96
5.8. Métricas promedio para la representación bidimensional con cinco clases y 40 coeficientes MFCC	98
5.9. Comparación entre representaciones 1D y 2D para el experimento con tres hablantes	99
5.10. Comparación entre representaciones 1D y 2D para el experimento con cinco hablantes y 100 MFCC	100
5.11. Comparación entre representaciones 1D y 2D para el experimento con cinco hablantes y 40 MFCC	101

Capítulo 1

Introducción

Con el paso del tiempo, las computadoras y, en general, la tecnología se utilizan en una gran variedad de campos. Prácticamente todo lo que involucre un proceso, de alguna manera, va acompañado de un sistema computacional. Bajo esta perspectiva, la comunicación puede considerarse también como un proceso susceptible de ser estudiado mediante herramientas computacionales. La comunicación es una parte esencial de muchos campos, como lo pueden ser la psicología, la sociología o incluso la antropología. El habla contiene información específica que va más allá del contenido puramente lingüístico. En ella se manifiestan emociones, actitud y tono, elementos que hacen que nuestra voz sea única y distinguible de todas las demás. Estas cualidades también son empleadas por los sistemas computacionales, que buscan convertirlas en datos y señales que puedan comprender. De esta manera, es posible interpretar y distinguir tales matices mediante el procesamiento de señales de audio.

Los avances en inteligencia artificial y el aprendizaje automático han dotado a los sistemas de una capacidad de procesamiento artificial suficiente para realizar tareas complejas, como interpretar patrones de datos del mundo real, incluyendo el audio, es decir, la voz humana. Específicamente, el procesamiento del lenguaje natural (PLN) es una subrama de la inteligencia artificial que utiliza técnicas de aprendizaje automático y aprendizaje profundo, y se encarga de estudiar el lenguaje natural en distintos niveles lingüísticos, con la finalidad de lograr un procesamiento similar al de los humanos. Como mencionan Belinkov y Glass (2018), en años recientes ha tenido progresos muy importantes con el avance de las redes neuronales. El reconocimiento del habla deriva del PLN y se enfoca en el estudio de las señales de voz, con el objetivo de identificar métricas tales como el hablante, palabras, oraciones, estado de ánimo, velocidad del habla, entre otras.

Dado este contexto general, el presente trabajo se enfoca en analizar el reconocimiento de hablantes a partir de la representación de la señal de voz mediante coeficientes cepstrales en la

escala de Mel (MFCC, por sus siglas en inglés: *Mel Frequency Cepstral Coefficients*). Se estudia el comportamiento de dos formas de representación distintas, lo cual se explicará más a fondo en la próxima sección.

1.1. Planteamiento del problema

En materia de la inteligencia artificial se encuentra el campo del procesamiento de audio, particularmente las tareas de reconocimiento y clasificación de hablantes. En estas tareas, los MFCC son elementos imprescindibles, pues permiten construir una representación compacta de las características de la voz y capturar información relevante de la estructura del habla. El uso de los MFCC normalmente se limita a su representación unidimensional (1D), es decir, a series numéricas, pues esta representación es históricamente eficaz cuando se combina con clasificadores supervisados como las máquinas de vectores de soporte (SVM, por sus siglas en inglés: *Support Vector Machines*) o los árboles de decisión, entre otros modelos de aprendizaje automático supervisado clásicos.

Con el surgimiento y la creciente adopción de las redes neuronales convolucionales (CNN, por sus siglas en inglés: *Convolutional Neural Networks*), diseñadas para el tratamiento de imágenes, ha surgido una nueva línea de investigación para aprovechar los MFCC. La interrogante en sí misma es acerca de la representación de estos datos, y en específico, si convertir las matrices de MFCC a imágenes y subsecuentemente tratarlas como datos visuales genera resultados satisfactorios, que puedan compararse o incluso sobrepasar aquellos obtenidos por representación en 1D. Esta transformación permite que las CNN preentrenadas hagan uso directo de estas representaciones visuales.

Aunque diversos trabajos han explorado la transformación bidimensional (2D) de los MFCC para tareas de clasificación de audio, aún resulta pertinente analizar con mayor detalle su desempeño en el reconocimiento de hablantes bajo condiciones experimentales controladas y comparables. En particular, no todos los estudios profundizan en la comparación entre representaciones

unidimensionales y bidimensionales. Por ello, resulta necesario evaluar esta alternativa de manera estructurada, considerando tanto su potencial como sus limitaciones en escenarios prácticos.

Esta situación plantea una cuestión interesante en cuanto al diseño de sistemas de reconocimiento de hablantes: ¿Qué tan efectiva resulta esta representación visual en aplicaciones prácticas reales? Esta pregunta es relevante desde una perspectiva técnica y práctica, ya que influye directamente en la elección de herramientas y métodos utilizados en la construcción de sistemas de reconocimiento de voz, especialmente en términos de la selección de herramientas y recursos adecuados para procesar representaciones visuales de MFCC.

1.2. Justificación

El reconocimiento del habla es una parte fundamental del procesamiento de señales de audio, y su importancia radica en sus aplicaciones, entre las cuales destacan la autenticación biométrica y los sistemas interactivos que funcionan mediante la voz, lo que requiere que un sistema sea capaz de identificar de manera precisa a la persona que emite la señal de voz.

Como mencionan Mahmood y Köse (2021), una de las técnicas más utilizadas en el área del reconocimiento del habla es el uso de los MFCC. Estos coeficientes reducen la complejidad del proceso al transformar el audio en un conjunto de vectores numéricos, lo que permite al sistema informático procesar una cantidad significativamente menor de datos en comparación con el análisis de la señal de audio en su forma original.

Por lo general, la aplicación de los MFCC para tareas de reconocimiento de hablantes se hace mediante su forma numérica unidimensional (1D), es decir, una sucesión de números organizada como vectores que representan los coeficientes extraídos de la señal de audio. Esta representación resulta efectiva, ya que por medio de ella se forma una base sólida para desarrollar sistemas de reconocimiento de voz y clasificación de hablantes. Sin embargo, el auge de las redes neuronales profundas y los avances tecnológicos relacionados con la visión por computadora generan una nueva fuente de posibilidades, entre las que se encuentra el interés de convertir estas sucesiones

numéricas en una forma visual, dicho de otro modo, generar imágenes a partir de los propios datos de los MFCC, que pudiesen ser interpretados por modelos diseñados originalmente para clasificar imágenes.

Este nuevo paradigma ha impulsado una creciente adopción de representaciones visuales en las tareas de clasificación de audio. En lugar de reducir los MFCC a un vector de números, se conserva su estructura completa en dos dimensiones. Al hacer esto, el audio se convierte en una matriz matemática que equivale a una imagen. Esto abre la puerta a usar redes neuronales convolucionales, que son herramientas diseñadas originalmente para la visión por computadora y la clasificación de imágenes (Jin et al., 2021).

A partir de este escenario, la presente investigación se justifica en la necesidad de realizar un estudio estructurado sobre el uso y desempeño de los MFCC en el contexto específico de la clasificación de hablantes. El interés principal es evaluar objetivamente esta metodología mediante el uso de datos reales para aportar evidencia clara sobre la efectividad y viabilidad del uso práctico de estas representaciones visuales, aclarando sus fortalezas y posibles limitaciones, y ofreciendo así un sustento sólido para su aplicación en sistemas reales.

Este estudio adquiere importancia debido a la sensibilidad que viene implícita en la clasificación de hablantes, donde incluso pequeñas mejoras en rendimiento pueden tener un impacto significativo en aplicaciones prácticas. Por tanto, la validación de estas formas de representación contribuye directamente a la mejora en la construcción de sistemas actuales de reconocimiento del habla.

1.2.1. Objetivo general

Desarrollar modelos computacionales para el reconocimiento de hablantes a partir de representaciones unidimensionales y bidimensionales de la señal de voz, con el fin de determinar cuál ofrece mejor rendimiento bajo un marco experimental comparable.

1.2.2. Objetivos específicos

1. Caracterizar la señal de voz en formato unidimensional para generar los vectores de entrada orientados a los algoritmos de clasificación supervisada.
2. Desarrollar modelos de clasificación supervisada para el reconocimiento de hablantes mediante el uso de representaciones unidimensionales.
3. Modelar la señal de voz en formato bidimensional, transformando los coeficientes en representaciones visuales con la finalidad de desarrollar modelos alimentados por imágenes.
4. Implementar y entrenar redes neuronales convolucionales para clasificar las representaciones bidimensionales de la voz, utilizando arquitecturas profundas preentrenadas y técnicas de ajuste de hiperparámetros.
5. Analizar el desempeño de los clasificadores que usan las representaciones generadas por ambos enfoques, para identificar ventajas y limitaciones de cada representación en la tarea de reconocimiento de hablantes.

1.3. Solución propuesta

La presente investigación plantea como solución la exploración y comparación de distintas formas de representar la señal de voz con el propósito de analizar y evaluar su impacto en la tarea de reconocimiento de hablantes. En particular, se consideran dos enfoques principales: representaciones unidimensionales, que organizan los coeficientes en vectores numéricos, y representaciones bidimensionales, que los disponen en formato visual.

Sobre estas representaciones se aplicarán diferentes modelos de clasificación, seleccionados de acuerdo con la naturaleza de los datos y ajustados mediante estrategias de optimización de parámetros. El proceso incluirá la búsqueda de conjuntos de datos adecuados, el entrenamiento de los modelos y la evaluación de su rendimiento bajo un mismo marco experimental.

Finalmente, los resultados obtenidos permitirán analizar las ventajas y limitaciones de cada enfoque, así como determinar si ambas representaciones pueden considerarse equivalentes en desempeño o si alguna ofrece un beneficio particular en determinados contextos de aplicación.

1.4. Organización de la tesis

Con el propósito de facilitar la lectura y el seguimiento de la investigación, el presente documento se encuentra distribuido en los siguientes capítulos principales:

En el Capítulo 1 se introduce el tema de estudio, se plantea el problema principal, se justifica la importancia de la investigación junto con los objetivos establecidos. y se describe de forma general la solución propuesta.

En el Capítulo 2 se detalla el marco conceptual que sostiene el proyecto. En este apartado se explican las características de la señal de voz, el procesamiento digital de los sonidos, el funcionamiento de los coeficientes MFCC y se realiza un recorrido sencillo por los métodos de aprendizaje automático y aprendizaje profundo necesarios para la clasificación.

En el Capítulo 3 se expone el estado del arte mediante una revisión de trabajos previos basados en formatos de representación unidimensionales y bidimensionales. De igual manera, se describen los conjuntos de datos de referencia recopilados para realizar el estudio junto con una discusión general de los antecedentes.

En el Capítulo 4 se muestra la metodología exacta diseñada para la investigación. En esta sección se detallan las etapas de extracción de características y clasificación, abarcando desde la preparación de los audios hasta la estructuración de la estrategia experimental.

En el Capítulo 5 se presentan los experimentos realizados y se analizan los resultados obtenidos. Aquí se evalúan de forma individual las pruebas ejecutadas con representaciones unidimensionales y bidimensionales, concluyendo con una comparación directa entre ambos enfoques y el análisis de los hallazgos clave.

Por último, en el Capítulo 6 se exponen las conclusiones generales derivadas de toda la

investigación. En este cierre se exponen los puntos más importantes del trabajo y se proponen posibles enfoques para el trabajo futuro.

Capítulo 2

Marco conceptual

En esta sección se presentan los conceptos teóricos necesarios para la investigación. En particular, se abordan desde los principios básicos de la voz humana hasta las representaciones utilizadas para lograr este fin.

2.1. La señal de voz

La voz humana es la herramienta principal de comunicación entre las personas y posee características únicas que permiten identificar a cada individuo. Al hablar, el aire que se expulsa genera ciertos sonidos que viajan por el entorno, transmitiendo el mensaje y ciertos detalles adicionales, tales como una gran cantidad de información sobre el estado físico, la identidad y las emociones del hablante.

The Voice Foundation (2019) indica que la voz es el producto de la combinación del sonido sonoro, la resonancia y la articulación:

- **Sonido sonoro:** El sonido básico que se produce por la vibración de las cuerdas vocales. Esta frecuencia produce una especie de sonido de *zumbido*. El sonido sonoro que se produce al cantar se diferencia mucho del producido en el habla.
- **Resonancia:** El sonido sonoro se amplifica o modifica por los resonadores del tracto vocal (la garganta, la cavidad bucal y las fosas nasales). Los resonadores son los responsables de que la voz de una persona sea tan reconocible y distinguible.
- **Articulación:** Los articuladores del tracto vocal (la lengua, el paladar suave y los labios) modifican el sonido de nuestra voz. Gracias a los articuladores podemos decir palabras que reconocemos.

Más allá de los factores fisiológicos que producen la voz, esta también se comporta como una señal de audio. Por lo tanto, para poder analizarla, es necesario estudiar sus propiedades desde el punto de vista acústico.

2.2. Características acústicas de la voz

La voz humana es una señal acústica que contiene una serie de propiedades y características tanto físicas como perceptibles que la hacen única para cada persona. Sin embargo, como mencionan Martínez Mascorro y Aguilar Torres (2013), estas características son variables tanto de forma voluntaria como involuntaria. Por lo tanto, comprender estas características es un paso esencial para el desarrollo de un sistema que logre interpretar y analizar las señales de voz de forma precisa sin caer en la confusión. Algunas de las propiedades clave de la voz se describen a continuación.

2.2.1. Tono

El tono representa la velocidad de vibración de las cuerdas vocales, y es el componente principal en la entonación. En muchas lenguas tiene una importancia clave, debido a que la entonación permite distinguir dos palabras que por lo demás serían idénticas. El tono es un fenómeno cuantificable que se emplea para describir las propiedades de la onda sonora (Akerberg et al., 2015).

2.2.2. Timbre

El timbre es la propiedad que permite distinguir dos voces que presentan el mismo tono e intensidad. En el ámbito musical, es la cualidad propia de un instrumento que logra diferenciarlo de otros, lo cual se manifiesta de la misma manera en la voz humana. Por ejemplo, gracias al timbre es posible identificar la diferencia entre dos personas que cantan exactamente la misma nota, ya que esta característica está relacionada con los matices únicos de cada sonido (Olguín Meza, 2023).

2.2.3. Intensidad

La intensidad está determinada directamente por la amplitud de la señal sonora, representa la fuerza o volumen con la que se percibe el sonido y depende principalmente de la presión acústica (Akerberg et al., 2015; Olguín Meza, 2023).

2.2.4. Duración

La duración se refiere al tiempo en que persiste la señal acústica sin discontinuidad. También se conoce como longitud, y se mide en milésimas de segundo (ms) (Akerberg et al., 2015; Olguín Meza, 2023).

2.2.5. Formantes

Los formantes son armónicos reforzados, *picos* de energía en el espectro de frecuencia de la voz, y son causados por la resonancia del tracto vocal. En los espectrogramas pueden apreciarse como bandas más oscuras (Akerberg et al., 2015).

2.3. Señal analógica y digital

Para que una computadora pueda entender lo que nosotros percibimos como la voz humana, es necesario representarla como una señal. Según Alvarado Erazo (2016), una señal es una portadora de información sobre una medición en particular, y puede ser discreta o continua en el tiempo.

Bajo este mismo enfoque, una señal analógica o natural puede presentarse de forma continua o discreta en el tiempo. Como se representa en la Figura 2.1, una señal es continua cuando sus valores varían de manera progresiva y sin interrupciones. Un ejemplo de este comportamiento es el voltaje de una batería, el cual cambia poco a poco pero sin saltos abruptos entre valores discretos.

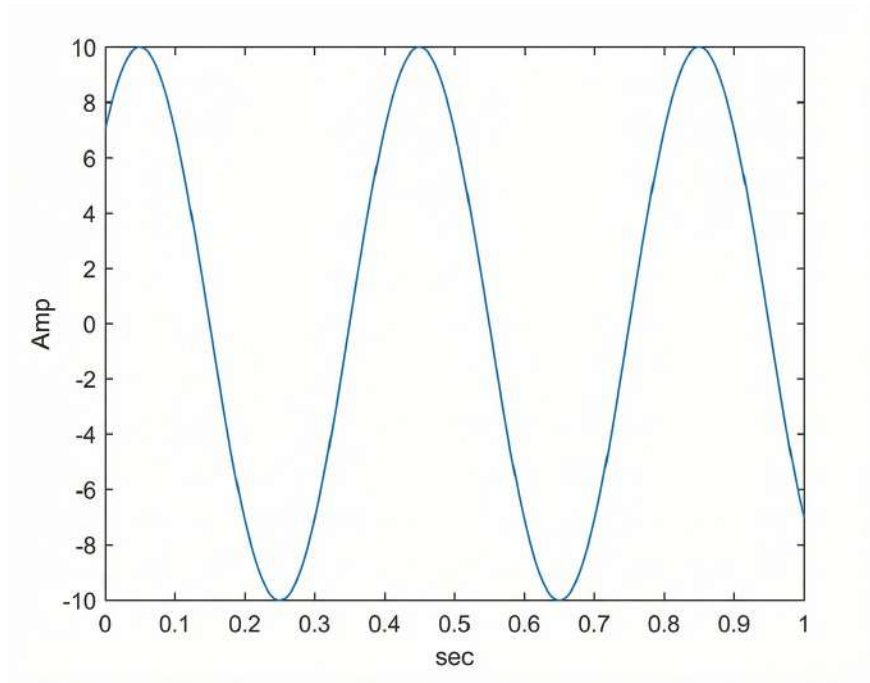


Figura 2.1. Ejemplo de una señal analógica que varía de forma continua a lo largo del tiempo.
Obtenido de Chamba y Jiménez (2022).

Por otra parte, una señal digital es aquella que solo toma valores discretos tanto en el tiempo como en amplitud, limitándose a valores finitos. La señal digital se convirtió en el estándar de la industria, ya que suele ser más exacta que la variante analógica debido a que es menos susceptible a interferencias externas (Alvarado Erazo, 2016). Este tipo de señal se ejemplifica en la Figura 2.2.

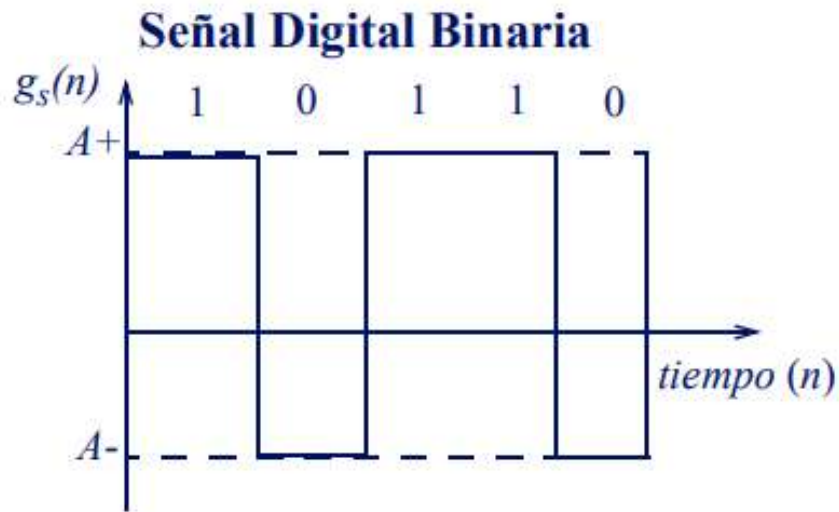


Figura 2.2. Representación gráfica de una señal digital binaria, en la que los valores pasan de 0 a 1 y viceversa. Obtenido de Departamento de Dispositivos Electrónicos (2003).

Comúnmente, para que una señal analógica pueda ser procesada por un sistema computacional, esta debe convertirse en una señal digital mediante procesos como el muestreo y la cuantificación. Como resultado, la información continua se representa mediante una secuencia discreta de valores numéricos. Un ejemplo de este tipo de representación se muestra en la Figura 2.3.

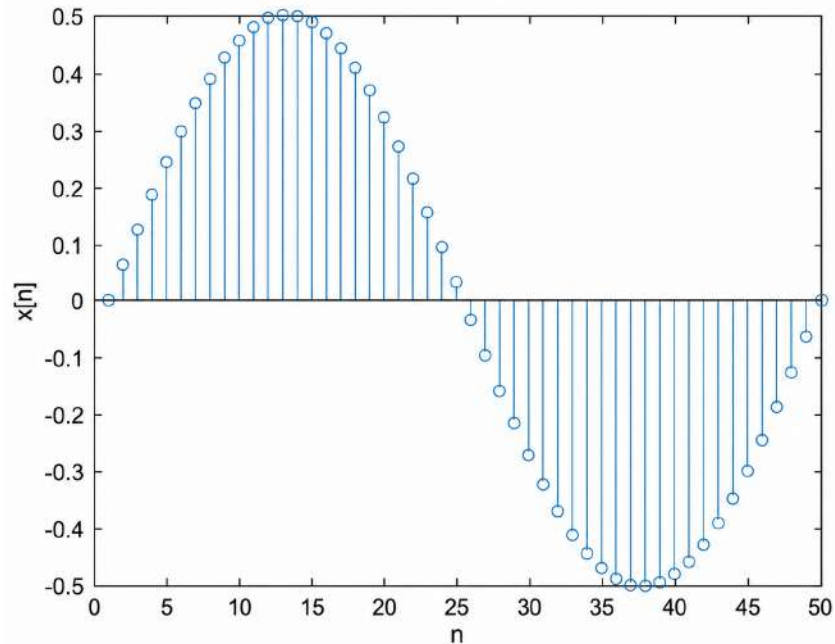


Figura 2.3. Representación de una señal discreta ($x[n]$), donde cada valor de la secuencia corresponde a una muestra definida para un índice entero (n). Adaptado de Chamba y Jiménez (2022).

El proceso de conversión de una señal analógica a una digital implica un posible riesgo de alterar la información original al adaptarla a un límite establecido. No obstante, para evitar pérdidas de información considerables, la señal analógica debe ser digitalizada con una tasa de muestreo suficientemente alta (Alvarado Moya, 2011).

A pesar de las restricciones del muestreo, es preferible utilizar un sistema digital sobre su equivalente analógico, ya que los sistemas digitales cuentan con algunas ventajas: son usualmente más baratos, confiables y ofrecen una precisión mucho mayor en el procesamiento de señales, evitando que el ruido vaya incrementándose de forma acumulativa en cada etapa de procesamiento. Además, al tratarse de sistemas programables, proveen una gran flexibilidad en el diseño, ya que sus características se pueden modificar o adaptar directamente a través de cambios en el software.

2.3.1. Importancia en el procesamiento de voz

La representación digital es necesaria para aplicar técnicas avanzadas de análisis y clasificación de hablantes. Esta transición de una señal continua a una secuencia discreta simplifica

significativamente la extracción de parámetros acústicos críticos, tales como los coeficientes ceps-
trales en la escala de Mel (MFCC, por sus siglas en inglés: *Mel Frequency Cepstral Coefficients*)
(Martínez Mascorro & Aguilar Torres, 2013), los cuales sirven posteriormente como la entrada
esencial para optimizar los modelos de aprendizaje automático y aprendizaje profundo.

2.4. Procesamiento digital de señales de voz

El procesamiento digital de señales de audio consiste en un conjunto de técnicas orientadas
a representar, analizar y manipular el sonido en formato numérico (series numéricas).

Oppenheim y Schafer (2011) mencionan que el procesamiento de señales se enfoca en
representar, manipular y transformar las señales junto con toda la información que estas contie-
nen. Esta disciplina permite, por ejemplo, realizar la separación de dos señales unidas o extraer
algún parámetro relevante de la señal. Además, Alvarado Erazo (2016) define el procesamiento de
señales como un proceso para obtener información deseada de una señal, como puede ser mejorar
la calidad de una señal por medio de la eliminación de ruido o frecuencias no deseadas.

De manera general, para que este procesamiento se pueda llevar a cabo, se suele requerir
de acciones relacionadas con preparar las señales de forma adecuada. Esto implica procesos tales
como la modulación, compresión o acondicionamiento, lo que permite a los algoritmos trabajar
con ellas correctamente (Oppenheim & Schafer, 2011).

Los objetivos del procesamiento de señales dependen de la aplicación, pero generalmente
incluyen:

- La mejora de una señal, con la finalidad de facilitar su análisis.
- La transformación de la señal para transmitirla o almacenarla.
- La extracción de información, pues de ello depende el reconocimiento de patrones y la cla-
sificación de la señal.

En la antigüedad, este tipo de procesamiento era altamente limitado debido a las capaci-

dades tecnológicas de la época. Sin embargo, en tiempos modernos, la era digital nos permite, gracias a la vasta cantidad de nuevo hardware especializado (procesadores y chips), impulsar desarrollos teóricos como la transformada rápida de Fourier (FFT, por sus siglas en inglés: *Fast Fourier Transform*), entre otros (Oppenheim & Schafer, 2011).

La aplicación de este tratamiento de señales abarca una gran cantidad de áreas, desde el ocio, las comunicaciones, los estudios de astronomía, la medicina y la arqueología, hasta campos de índole militar, entre otros (Oppenheim & Schafer, 2011). En particular, cuando se trata de señales de audio, surge la necesidad de representarlas de alguna manera para observar la variación temporal de las ondas, además de permitir el análisis de sus componentes espectrales y las frecuencias que conforman la señal.

De acuerdo con González et al. (2019), quienes realizan una modificación orientada a la programación a partir del modelo de verificación del hablante de Charisma et al. (2017), se sugiere que una manera de estructurar y clasificar las etapas del procesamiento de señales de voz para reconocimiento del hablante es a través del siguiente flujo (véase Figura 2.4):

- Fuente, captura y digitalización: la señal analógica (la voz) se captura desde un micrófono o dispositivo de entrada y luego se transforma en una señal digital por medio del muestreo y la cuantificación.
- Preprocesamiento: consiste en eliminar o controlar todo lo que pudiese influir en la calidad de la señal, comúnmente conocido como ruido, la eliminación de silencios y segmentación para ubicar los segmentos donde se habla más activamente. Además, se aplica normalización y filtros que mejoren la condición de la señal.
- Extracción de características: se extraen parámetros que representen la información, como los MFCC, formantes, frecuencia fundamental y demás características que logren describir la señal de forma compacta y útil.
- Modelo de referencia y semejanza (sistema clasificador): bloque encargado de comparar los parámetros extraídos de la voz actual contra los patrones matemáticos de las voces que ya

están guardadas en la base de datos, midiendo qué tan parecidos son entre sí.

- Resultado de identificación (identificador de hablante): bloque de salida resultante que procesa las métricas del clasificador para concluir la identidad del hablante evaluado.

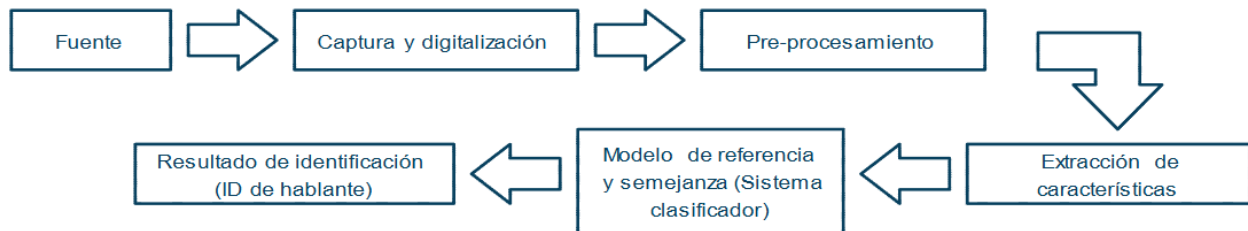


Figura 2.4. Diagrama de las etapas en un sistema de reconocimiento del hablante. Obtenido de González et al. (2019).

2.4.1. Dificultades y desafíos en el procesamiento digital de señales de voz

A pesar de los avances en las técnicas digitales, el procesamiento de la voz continúa enfrentando retos importantes que influyen en la calidad de los sistemas de análisis y reconocimiento. Bishop (1995) explica que, aunque para los seres humanos pudieran parecer problemas que requieren poco esfuerzo, en muchos casos solucionarlos mediante procesamiento digital se vuelve extremadamente difícil para las computadoras.

De acuerdo con Moreno y Rüdinger (2024), los sistemas que trabajan con señales de voz se enfrentan a varios desafíos en condiciones reales, los cuales van desde el ruido del entorno hasta problemas de seguridad. Con base en lo descrito por dicho autor, algunos de los retos principales se resumen en los siguientes puntos:

- Uso en entornos no controlados y ruido ambiental: Cuando el usuario interactúa con el sistema en escenarios cotidianos, la señal de voz se mezcla con sonidos de fondo, interrupciones y distorsiones. Esto hace que sea necesario implementar etapas de preprocesamiento para limpiar el audio y rescatar la señal de voz.

- de la voz: La voz de una misma persona cambia constantemente y presenta diferentes formas fonéticas. El reto consiste en lograr que la etapa de extracción de características obtenga parámetros que representen la información del hablante de forma útil y estable, logrando sobrepasar estas variaciones.
- Falsificación de voz y seguridad: En sistemas de reconocimiento del hablante, existe el riesgo de que se intente falsificar la voz para acceder a información confidencial. Para evitarlo, el sistema debe ser preciso al evaluar las métricas para dar un resultado de identificación correcto.
- Privacidad de los datos en la nube: Procesar o almacenar los audios en servicios web externos genera riesgos de privacidad, debido a las filtraciones de audio que han sufrido empresas de asistentes inteligentes.

2.4.2. Frecuencia y percepción auditiva

Según la Real Academia Española (2026), la definición de frecuencia es “número de veces que se repite un proceso periódico por unidad de tiempo”. En el contexto de la señal de la voz humana, este parámetro determina el tono percibido, donde las frecuencias bajas corresponden a voces graves y las altas se asocian con voces agudas.

A mayor frecuencia, menos espacio existe entre las oscilaciones de la onda, lo que significa que estas ocurren de forma más rápida (véase la Figura 2.5). Para el análisis de señales de audio, en la literatura de Sintec Insonorización (2021) se sugiere una descomposición general del rango audible en tres bandas frecuenciales específicas: frecuencias bajas o graves para valores menores a 250 Hz, frecuencias medias entre 250 Hz y 1,000 Hz, y frecuencias altas o agudas para valores que superan los 1,000 Hz.

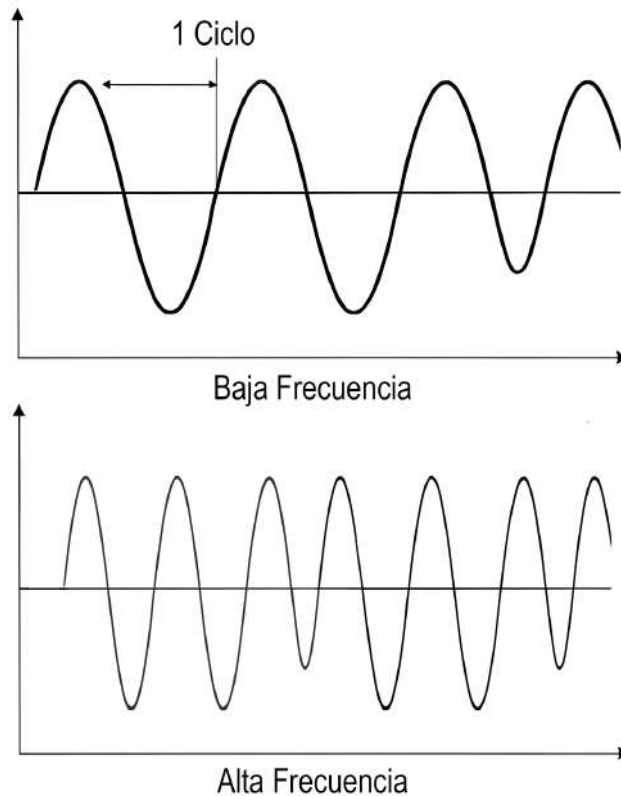


Figura 2.5. Representación de una frecuencia alta y baja.

Por otro lado, es fundamental saber cómo interactúa la frecuencia de las ondas de sonido con la percepción del oído humano, separando el volumen de un sonido de su tono, es decir, si es grave o agudo. De acuerdo con Sintec Insonorización (2021), el sistema auditivo no percibe todas las frecuencias con la misma intensidad. Como se ilustra en las curvas de respuesta de la Figura 2.6, nuestro oído es muy sensible a los sonidos que se encuentran en el rango de la voz humana (zona etiquetada como ‘comunicación’), pero es de alguna manera más ‘sordo’ a los sonidos muy graves. Esto significa que los tonos bajos necesitan reproducirse con un volumen mayor para que el oído pueda notarlos mejor.

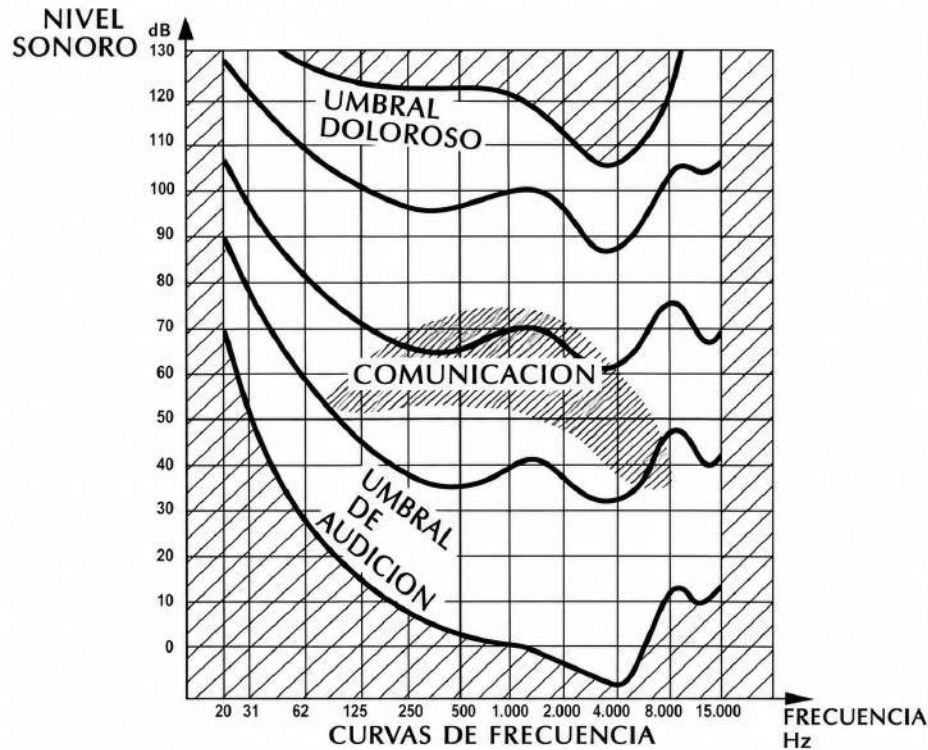


Figura 2.6. Curvas de respuesta en frecuencia, umbrales de audición y dolor del oído humano (Sintec Insonorización, 2021).

En cambio, esta respuesta en volumen contrasta con la capacidad del aparato auditivo para identificar el cambio en las notas musicales. En cuanto al tono, el oído humano muestra un comportamiento subjetivo y no lineal. Como explica González González (2019), si observamos el teclado de un piano (véase la Figura 2.7), para nuestro cerebro el salto de una tecla a otra se percibe siempre como una distancia idéntica, es decir, un intervalo musical llamado medio tono. Sin embargo, en el mundo físico de las frecuencias, la distancia real en hercios (Hz) se va estirando a medida que subimos hacia los agudos.

Para entender esto de forma más clara, González González (2019) expone el comportamiento de tres notas consecutivas en el piano:

- Al partir de la nota La (A_4), la cual vibra a una frecuencia fija de 440 Hz, y avanzar un tono completo hasta llegar a la nota Si (B_4), situada a 493.88 Hz, la diferencia física real entre ambas notas es de 53.88 Hz.

- Al avanzar exactamente un tono más desde dicha nota Si (B_4) para llegar a la nota Do sostenido ($C_5^\sharp$), la cual se ubica en los 554.37 Hz, la diferencia física resultante es de 60.49 Hz.

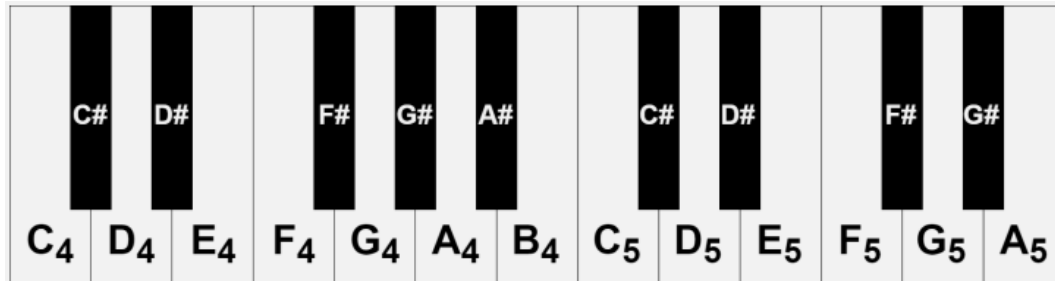


Figura 2.7. Representación de la escala musical en el sistema anglosajón donde la distancia entre cada tecla adyacente es de medio tono (Adaptado con fines ilustrativos a partir del ejemplo didáctico propuesto por González González, 2019).

Matemáticamente, el segundo salto es un 12.26 % más grande en frecuencia que el primero. Sin embargo, al escuchar ambas transiciones en el piano, el cerebro ignora esa diferencia numérica y percibe exactamente la misma distancia tonal entre los dos pares de notas.

Esto demuestra que el oído humano es sumamente preciso detectando variaciones muy pequeñas de frecuencia en los sonidos graves, pero se torna menos detallista mientras avanza en los sonidos agudos, procesando la información de manera logarítmica y no lineal. Este comportamiento representa un desafío claro si se desea que una computadora analice la voz de la misma manera en que lo hace un ser humano.

Precisamente, debido a esta naturaleza no lineal de la percepción tonal, el análisis computacional del habla requiere metodologías capaces de extraer las propiedades más relevantes de la señal basándose en el funcionamiento biológico de la audición. Después de un procesamiento inicial, estas propiedades de la muestra de habla, las cuales pueden ser tanto acústicas como fonéticas, se obtienen mediante diversos métodos de extracción de características, entre los que destacan la predicción lineal de la voz (LPC, por sus siglas en inglés: *Linear Predictive Coding*), el análisis relativo espectral (RASTA, por sus siglas en inglés: *Relative Spectral Transform*) o los coeficientes cepstrales en la escala de Mel (MFCC, por sus siglas en inglés: *Mel-Frequency Cepstral Coeffi-*

cients). No obstante, en el presente trabajo se ahondará con mayor profundidad en una técnica basada en MFCC.

2.4.3. Escala de Mel

Como se abordó en la sección anterior, diversos estudios señalan que los seres humanos percibimos las frecuencias de una forma subjetiva y no lineal. En este contexto, Stevens et al. (1937) establecieron una escala que buscaba adecuarse a la audición real, demostrando desde un punto de vista psicoacústico que la forma en que interpretamos la altura de los sonidos difiere de la linealidad matemática de la frecuencia física.

El experimento consistió en solicitar a cinco voluntarios que escucharan una frecuencia de referencia para luego ajustar una perilla hasta el punto en que el segundo tono fuera percibido como la mitad de agudo que el primero. Los resultados evidenciaron la irregularidad de la percepción humana a medida que aumenta la altura tonal, ya que, en las frecuencias más altas, el sistema auditivo requería una reducción en hercios notablemente más drástica para detectar formalmente que el tono había bajado a la mitad (véase la Figura 2.8).

FRECUENCIA ESTÁNDAR	FRECUENCIA A MEDIO TONO						ERROR PROMEDIO (%)	
	OBSERVADOR					MEDIA GEOMÉTRICA	OBSERVADOR	
	<i>E</i>	<i>V</i>	<i>N</i>	<i>M</i>	<i>D</i>		<i>E</i>	<i>V</i>
125	96	84	97	87	96	90	3.4	7.5
200	131	131	123	101	107	121	7.4	6.7
300	183	205	195	118	133	171	13.0	7.8
400	237	251	237	173	253	233	13.5	9.0
700	391	407	474	256	403	384	12.3	9.8
1,000	590	640	622	391	485	558	11.0	13.9
2,000	998	979	1006	508	662	851	14.8	13.7
5,000	1930	1800	2260	1230	1590	1767	12.0	13.5
8,000	2360	2170	3110	1650	2110	2239	19.1	13.9
12,000	2970	2400	3720	2320	2980	2788	19.9	11.5

Figura 2.8. Resultados de los cinco participantes del experimento, mostrando sus medias geométricas y los errores promedio de dos observadores. Adaptado (traducido) de Stevens et al. (1937).

A partir de estos resultados, los autores graficaron y trazaron la curva empírica que definiría la escala de Mel. Como se aprecia en la Figura 2.9, esta representación transforma los hercios físicos en unidades mel perceptuales, adaptándose con mayor fidelidad a la capacidad real del aparato auditivo para procesar el espectro sonoro.

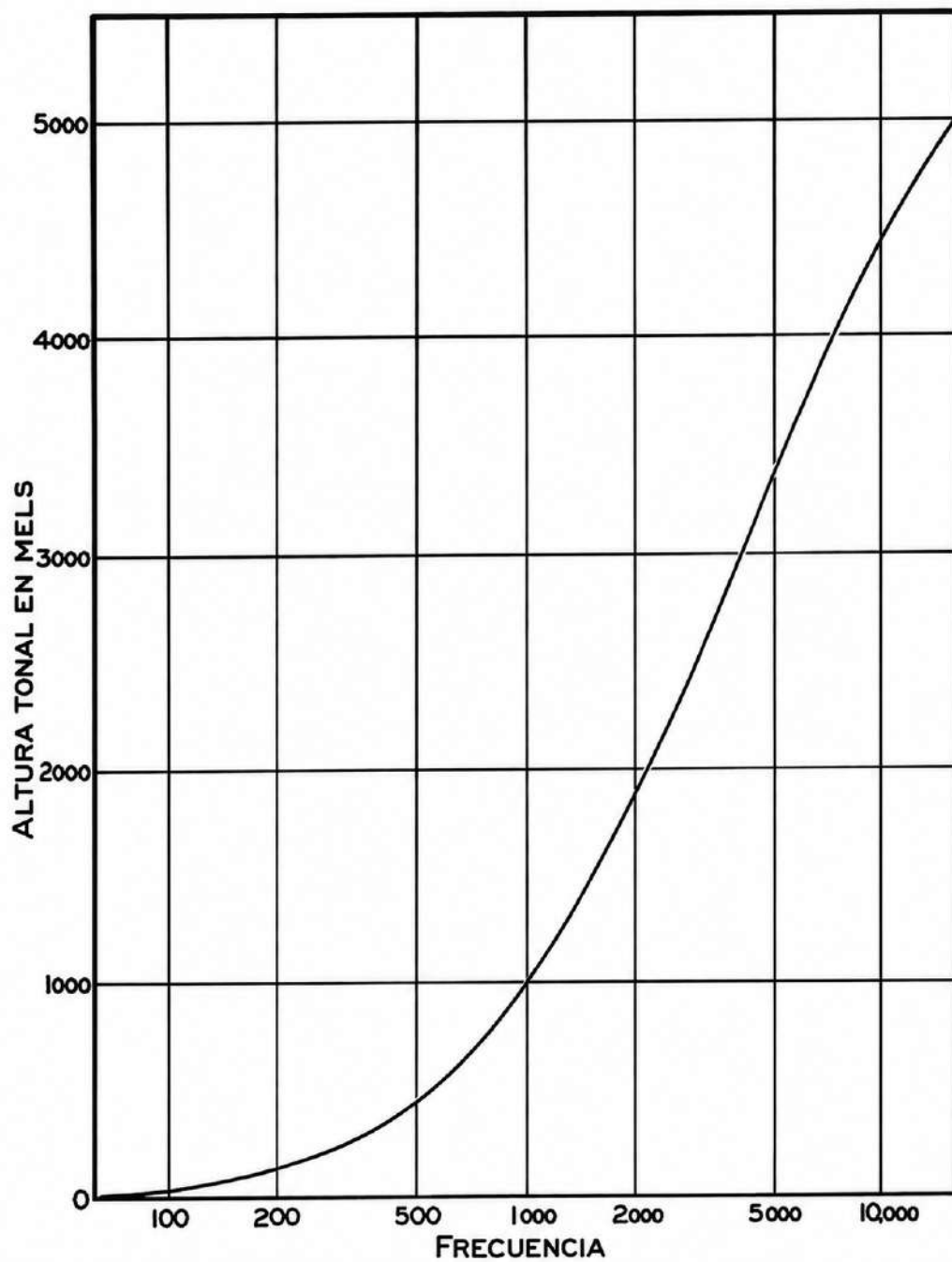


Figura 2.9. Gráfica original de la escala Mel que relaciona la frecuencia física (Hz) con la altura tonal percibida (mels). Adaptado (traducido) de Stevens et al. (1937).

Esta correspondencia entre la frecuencia física y la interpretación subjetiva del cerebro constituye el pilar fundamental para el procesamiento digital de señales de voz adaptado al ser humano. En lugar de analizar el habla de manera matemática y lineal, los sistemas computacionales aprovechan esta compresión logarítmica para ignorar las variaciones irrelevantes en las zonas agudas y priorizar el detalle fino en los rangos graves. Es bajo este principio donde encuentran su justificación los MFCC, que se describirán a continuación.

2.5. Coeficientes cepstrales en la escala de Mel (MFCC)

Los MFCC están basados en la escala de Mel, lo que significa que modelan la forma en que los seres humanos perciben el sonido, proporcionando una representación compacta, pero al mismo tiempo robusta, de la señal de voz. Al emplear estos coeficientes, se logra capturar las características espectrales más relevantes de cada audio, lo que permite una mejor clasificación debido a que representan tanto la envolvente del espectro como su fluctuación en el dominio del tiempo (Vázquez Robledo et al., 2024). Básicamente, los MFCC se utilizan para describir el timbre del sonido, por lo que son ampliamente empleados tanto para el reconocimiento del hablante como para la identificación de instrumentos musicales (Nagawade & Ratnaparkhe, 2017).

En un nivel práctico, los MFCC ofrecen un equilibrio óptimo entre simplicidad computacional y capacidad de discriminación. Al emplear un número relativamente pequeño de coeficientes (comúnmente entre 15 y 39 por ventana de análisis (Camargo Abril, 2024)), es posible representar adecuadamente las características más relevantes de la voz. Debido a esto, este método se consolida como una de las técnicas de extracción de parámetros más exitosas en el reconocimiento automático del habla, permitiendo condensar la señal acústica en vectores manejables que resultan altamente efectivos para discriminar patrones acústicos y conservar los atributos esenciales requeridos en las etapas de clasificación (Hegde et al., 2015).

Ali et al. (2020) resumen las etapas del cálculo de los MFCC de la siguiente manera:

- **Preénfasis:** La señal de voz se hace pasar a través de un filtro de paso alto con el objetivo de

amplificar las frecuencias altas y compensar la caída espectral natural de la voz. Este filtrado incrementa la energía de los componentes agudos de la señal.

- Segmentación en marcos: La señal de habla continua se divide en pequeños segmentos temporales, comúnmente denominados marcos, con una duración aproximada de entre 20 y 30 milisegundos cada uno. En esta etapa se deben configurar parámetros críticos como la longitud del marco (N) y el desplazamiento o grado de solapamiento entre marcos adyacentes (M), tal como se ilustra en la Figura 2.10. La justificación técnica del solapamiento se complementa en la siguiente etapa correspondiente al ventaneo.

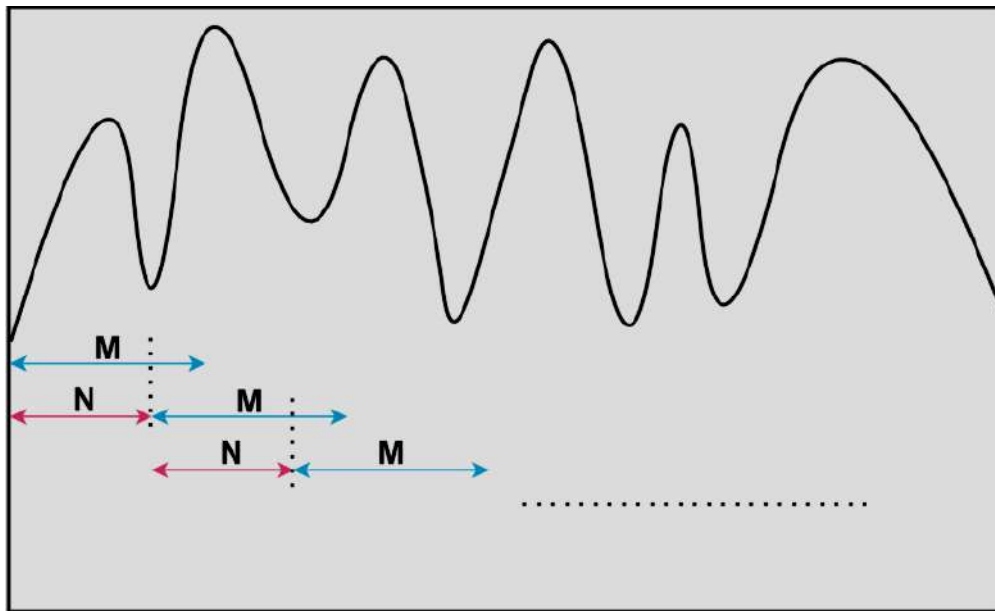


Figura 2.10. Ejemplificación de la segmentación en marcos de una señal de voz. Imagen obtenida de Yan et al. (2024).

- Ventaneo: Es una técnica común en el procesamiento de señales que consiste en aplicar una función suavizante a cada marco temporal. La función de ventaneo suaviza la señal a cero en los extremos. Por esta razón, es necesario el solapamiento de los marcos, ya que sin él la información clave en los extremos se perdería. Posterior a la aplicación de esta función, la señal cambia. Sin embargo, los efectos que tiene sobre las estadísticas globales son mínimos. La función de ventaneo más utilizada en conjunto con los MFCC es la ventana de Hamming,

la cual reduce la amplitud en los bordes, pero no la elimina por completo, lo que permite una transición más suave entre marcos solapados, como se aprecia en la Figura 2.11.

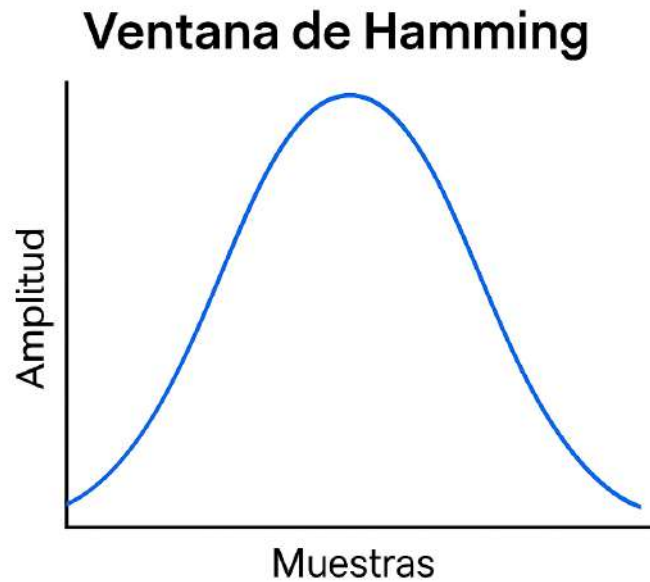


Figura 2.11. Representación geométrica de la función de ventana de Hamming.

- Transformada rápida de Fourier: Es una herramienta fundamental de análisis espectral que se utiliza para descomponer cada marco de tiempo en sus componentes de frecuencia sin perder información relevante. Mediante este paso, la señal se convierte del dominio del tiempo al dominio de la frecuencia (véase la Figura 2.12), facilitando la extracción posterior de las características acústicas.

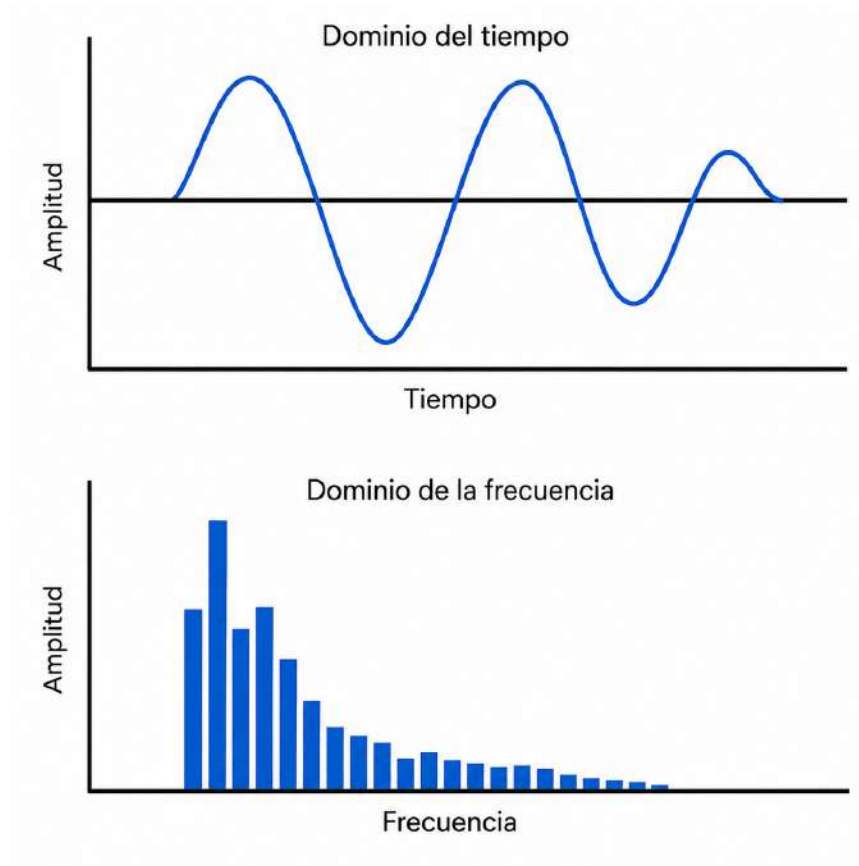


Figura 2.12. Representación de la señal en el dominio del tiempo y la frecuencia.

- Bancos de filtros en escala Mel: El oído distingue mejor las bajas frecuencias que las altas. Para imitar eso, el espectro se pasa por un conjunto de filtros triangulares distribuidos en la escala Mel. Así, se obtienen unas cuantas bandas de energía con más detalle donde importa más perceptualmente.
- Compresión logarítmica: Se toma el logaritmo de la energía de cada banda, lo que atenúa el efecto de picos muy elevados, acercándose a cómo percibimos los cambios de intensidad.
- Transformada discreta del coseno (DCT, por sus siglas en inglés: *Discrete Cosine Transform*): Debido a que las bandas de frecuencia de la escala Mel se encuentran fuertemente vinculadas entre sí, se aplica la DCT para separar matemáticamente sus componentes y eliminar la redundancia. Este paso transforma la señal de regreso a un dominio temporal modificado, llamado dominio cepstral, permitiendo concentrar la información más importante

del habla en un vector compuesto únicamente por los primeros 12 o 13 coeficientes.

Aunque los MFCC son altamente efectivos para representar la señal en marcos temporales cortos, de forma aislada solo capturan la información estática dentro de cada fragmento y no reflejan las transiciones dinámicas entre marcos consecutivos. Este aspecto se resuelve mediante la incorporación de los denominados coeficientes *delta* y *delta-delta*. Estos parámetros dinámicos enriquecen la representación acústica y se describen a continuación.

2.5.1. Delta y delta-delta

Junto con el cálculo de los MFCC de forma convencional, también es común que se obtengan sus derivadas temporales, conocidas como delta y delta-delta. Esto se debe a la naturaleza estática de los MFCC, los cuales describen cómo suena la voz en períodos cortos, pero no capturan la dinámica temporal de los coeficientes. Incluir estas derivadas junto con los MFCC permite registrar dicha variación, ampliando así la representación con información sobre la evolución del habla a lo largo del tiempo.

Estos parámetros se describen a continuación:

- Delta: Denominados coeficientes de velocidad o primera derivada, representan la variación de los coeficientes MFCC respecto a un lapso de tiempo (Gareta Manzano, 2015).
- Delta-delta: También llamados coeficientes de aceleración, se obtienen mediante la segunda derivada temporal de los MFCC. Los cambios temporales en el cepstro tienen un rol importante en la percepción humana, y es a través de estas derivadas que se logran medir dichos cambios (Nouhaila et al., 2022). Básicamente, estos coeficientes siguen el mismo proceso de cálculo que los delta, pero aplicados sobre estos últimos, midiendo la velocidad con la que cambian y justificando así su nombre de coeficientes de aceleración (Gareta Manzano, 2015).

Los coeficientes delta y delta-delta pueden mejorar el desempeño de los modelos al proveerlos de información adicional que en muchos casos resulta indispensable. Esto se debe a que la

voz de un individuo no solo está determinada por la forma de su tracto vocal (timbre), sino también por la transición dinámica de sus sonidos. En este sentido, mientras que los delta capturan las variaciones significativas en el tiempo, los delta-delta proporcionan información sobre la rapidez con la que están cambiando estas primeras características, lo cual resulta un aspecto crucial para tareas que dependen estrictamente del seguimiento de la dinámica temporal de la señal de audio (Camarago Abril, 2024). No obstante, existen situaciones en las que, más que aumentar la precisión, estos parámetros incrementan el riesgo de sobreajuste. Por ello, su inclusión no se considera obligatoria, sino que está directamente relacionada con el contexto y los objetivos específicos del problema.

2.6. Procesamiento del lenguaje natural (PLN) y su relación con el reconocimiento de hablantes

El procesamiento del lenguaje natural (PLN) es un área de la inteligencia artificial cuya tarea es buscar que las máquinas puedan comprender e interactuar con el lenguaje humano, tal como se describe en Stryker y Holdsworth (2024):

“El PLN permite que las computadoras y los dispositivos digitales reconozcan, comprendan y generen texto y voz mediante la combinación de la lingüística computacional (el modelado basado en reglas del lenguaje humano) y el modelado estadístico, el aprendizaje automático y el aprendizaje profundo.”

Bajo esta premisa, el PLN aporta las herramientas necesarias para interpretar el contenido lingüístico de un enunciado. Esto involucra técnicas capaces de convertir texto o voz en información estructurada, permitiendo así una comunicación eficiente entre el hombre y la máquina (Celi-Parraga et al., 2021).

La relación entre el procesamiento de voz y el PLN se da a través de los sistemas de reconocimiento automático del habla (ASR, por sus siglas en inglés: *Automatic Speech Recognition*), los cuales convierten la señal de audio en texto. A partir de esta conversión, el PLN aplica sus

técnicas para extraer el significado y el contexto del mensaje. Este flujo de trabajo permite combinar el análisis de la señal con el del texto para apoyar tareas como la identificación de hablantes en diálogos y el análisis emocional basado en lo que se dice.

2.6.1. Reconocimiento del hablante

El reconocimiento de hablantes es un área dentro del procesamiento de voz cuya finalidad es detectar la identidad del individuo que está hablando, mediante la identificación de las características únicas de su voz.

Tirumala et al. (2017) mencionan que el reconocimiento del hablante es el proceso de identificar a un individuo según las características vocales de una muestra de habla determinada. Esto difiere del reconocimiento del habla, debido a que este último se centra en descifrar el contenido del mensaje y no en el emisor. El proceso se basa en identificar y extraer los rasgos únicos de la voz de una persona, concepto también conocido como biometría de la voz. De esta manera, un sistema de reconocimiento de hablantes se emplea para extraer características particulares que permiten la verificación o autenticación del usuario.

Por otro lado, desde un punto de vista práctico, Microsoft Corporation (2025) se refiere al reconocimiento del hablante como un conjunto de algoritmos que identifican y verifican la identidad de las personas por medio de perfiles de voz únicos, utilizando los datos biométricos de audio para crear firmas de voz que puedan ser comparadas y validadas. Este proceso se puede observar en la Figura 2.13. Este flujo puede ser dependiente del texto o independiente del texto, variando en las restricciones que se tienen al momento de la inscripción. El proceso de identificación del hablante ayuda a determinar a un hablante desconocido dentro de un grupo de hablantes inscritos, y también permite atribuir voz a hablantes individuales para aprovechar el potencial de escenarios con varios hablantes, como los dispositivos multiusuario.

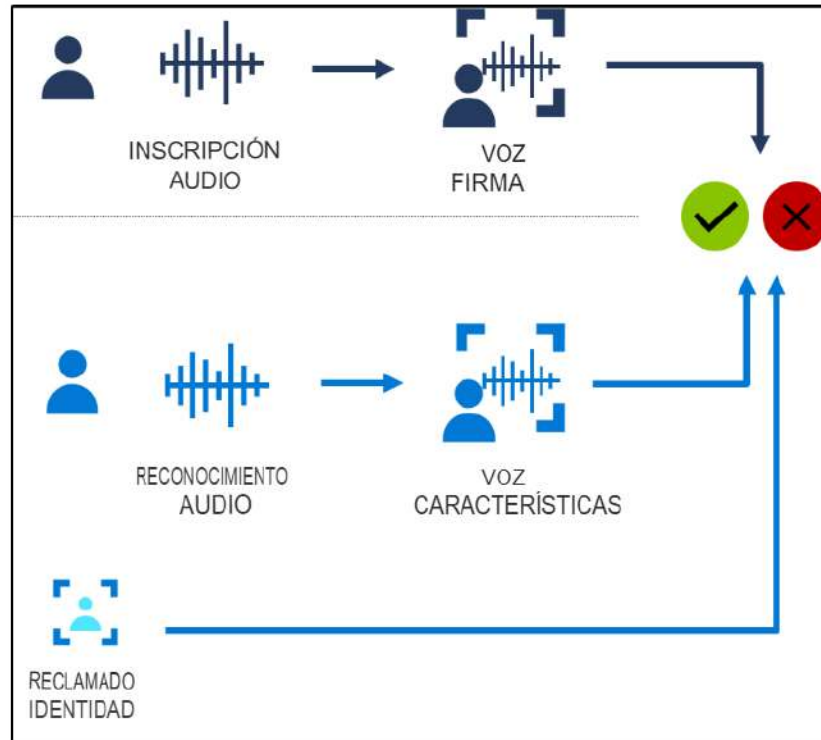


Figura 2.13. Diagrama de flujo de la verificación del hablante. Adaptado (traducido) de Microsoft Corporation (2025).

Aplicaciones prácticas

Mohamed et al. (2022) resumen algunas de las aplicaciones más importantes del reconocimiento de hablantes:

- **Autenticación:** consiste en verificar la identidad de una persona en sistemas sensibles para reemplazar o complementar las contraseñas tradicionales. Al basarse en la biometría de voz, permite un acceso rápido, seguro y sin fricciones. Suele aplicarse en la banca en línea, centros de atención telefónica (call centers), dispositivos móviles y sistemas de internet de las cosas (IoT, por sus siglas en inglés: *Internet of Things*). A pesar de sus ventajas, requiere mecanismos de protección contra fraudes mediante grabaciones o réplicas de voz artificiales.
- **Reconocimiento forense de hablantes:** en el ámbito legal, se busca comparar grabaciones de voz para determinar si pertenecen a una misma persona, lo que puede servir como eviden-

cia en investigaciones criminales y juicios legales. Factores como el ruido, la distorsión, la duración y la condición de las muestras son clave para la fiabilidad del análisis.

- Supervisión y vigilancia: en el contexto de la seguridad pública, se emplea para rastrear e identificar individuos específicos dentro de grandes volúmenes de grabaciones de audio. Más que tomar decisiones definitivas, estos sistemas actúan como herramientas de filtrado para priorizar muestras de interés que posteriormente serán analizadas por expertos, por lo que requieren algoritmos robustos capaces de operar con registros parciales.
- Seguimiento de hablantes: permite detectar los cambios de turno en una conversación continua para mantener la identidad de los participantes a lo largo del tiempo. Este proceso de organización y segmentación de datos de audio se conoce formalmente como diarización del hablante (*speaker diarization*) y se encuentra presente en la transcripción automatizada de reuniones virtuales, debates y sistemas de conferencia.
- Seguridad general: a diferencia de la autenticación individual, este enfoque se centra en proteger canales enteros de comunicación, como redes de llamadas por internet, contra accesos no autorizados o interceptaciones. Utilizar la identidad de la voz en la infraestructura de red ofrece una capa de protección masiva en escenarios donde los datos viajan cifrados y no se cuenta con cámaras para reconocimiento facial. El reto principal aquí consiste en garantizar la privacidad de los flujos de datos a gran escala.

2.6.2. Diferencia entre reconocimiento de hablante y de voz

Los conceptos de hablante y habla pudieran parecer referirse a lo mismo; sin embargo, son áreas diferentes dentro del procesamiento de voz. Reconocer al hablante tiene por objetivo determinar la identidad de la persona que está emitiendo el mensaje, mientras que reconocer el habla se centra en descifrar el contenido lingüístico y fonético de lo que se pronunció.

En el reconocimiento del habla, las variaciones individuales de la voz son tratadas como ruido que debe minimizarse para lograr una transcripción precisa del mensaje. Por el contrario, en

el reconocimiento del hablante esas variaciones son la fuente de información más valiosa, ya que permiten distinguir entre individuos. Esta diferencia en los objetivos conduce también a diferencias en los métodos empleados: los sistemas de reconocimiento del habla suelen priorizar modelos que abstraen el contenido y normalizan la diversidad de voces, mientras que los de reconocimiento del hablante buscan descriptores que resalten las particularidades acústicas de cada persona.

Por consiguiente, sus criterios de éxito también son distintos. Para el reconocimiento del habla, la calidad de un sistema se mide en función de la fidelidad de lo que se ha dicho respecto a lo que se detectó en la transcripción del mensaje. En cambio, en el reconocimiento del hablante, lo que importa es validar la identidad del emisor. Esto deja ver que, aunque pueden compartir etapas de procesamiento, los objetivos finales y la revisión de resultados son divergentes.

Si bien el objetivo central de este trabajo no es el análisis semántico del habla, la relación entre voz y PLN es relevante para ubicar la investigación dentro del ecosistema más amplio del procesamiento del lenguaje hablado. El reconocimiento de hablantes, aunque se centra en rasgos acústicos, puede integrarse en aplicaciones donde el PLN también está presente, como en asistentes inteligentes capaces de distinguir usuarios o en entornos de seguridad que requieren identificar no solo lo que se dice, sino también quién lo dice.

2.6.3. Representaciones de coeficientes MFCC usando espacios 1D y 2D

Una vez que los MFCC se calculan con éxito, estos se pueden organizar de diferentes maneras según el modelo que se vaya a utilizar. En este trabajo interesa comparar dos enfoques principales: usarlos como vectores unidimensionales (1D) o acomodarlos en forma de imagen (2D).

Representación 1D

En el enfoque unidimensional (1D), toda la información de un audio se convierte en un solo vector de tamaño fijo. Para lograr esto, se toma la matriz original de los MFCC y se “aplana” (*flattening*). Esto significa que los datos de cada fragmento de tiempo se colocan uno detrás de otro

en una sola línea continua. Así, toda la información del audio queda unificada en una sola lista, lo que facilita la entrega de los datos a los modelos de clasificación tradicionales.

- **Ventajas:** es una estructura muy simple, requiere menor almacenamiento y funciona directamente con algoritmos clásicos como máquinas de vectores de soporte (SVM, por sus siglas en inglés: *Support Vector Machines*), k vecinos más cercanos (k -NN, por sus siglas en inglés: *k-Nearest Neighbors*) o bosques aleatorios sin necesidad de realizar modificaciones.
- **Limitaciones:** al estirar la matriz y dejar todos los números en una sola fila, se pierde la estructura bidimensional original, lo que complica que el modelo identifique cómo se relacionan los fragmentos de tiempo entre sí de forma geométrica o visual.

Representación 2D

En contraste con el enfoque anterior, el enfoque bidimensional (2D) organiza los MFCC en forma de matrices o imágenes en lugar de filas de números. Los coeficientes se acomodan en una cuadrícula de tiempo contra coeficientes, de manera muy similar a un espectrograma, como se ilustra en la Figura 2.14. Esto produce una imagen digital donde un eje representa el transcurso del tiempo y el otro corresponde a los diferentes coeficientes calculados.

- **Ventajas:** conserva la evolución del sonido a lo largo del tiempo y permite utilizar redes neuronales convolucionales (CNN), las cuales son capaces de detectar patrones visuales directamente en la imagen de la voz.
- **Limitaciones:** requiere una mayor capacidad de cómputo, exige igualar el tamaño de los audios (mediante recortes o relleno de ceros) y necesita conjuntos de datos más grandes para evitar que el modelo se memorice los datos en exceso.

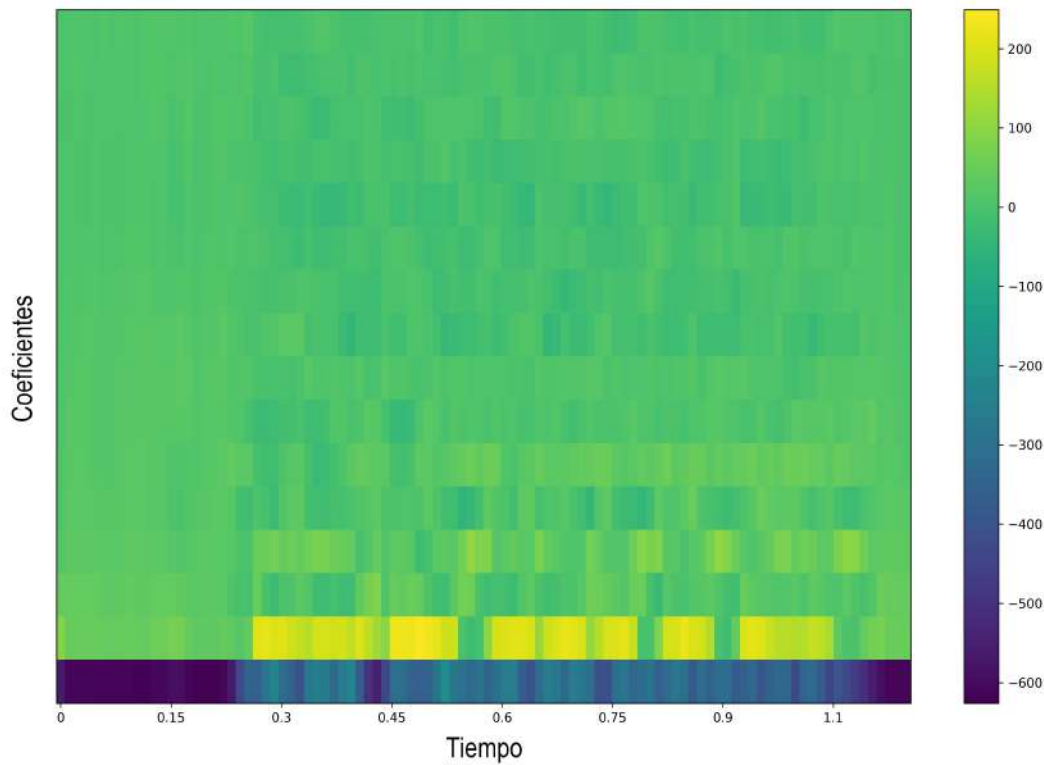


Figura 2.14. Ejemplo de representación 2D de MFCC.

De este modo, la representación 1D resulta más adecuada para conjuntos de datos pequeños al reducir el riesgo de sobreajuste. Por el contrario, la representación 2D es superior para capturar patrones de voz muy complejos, pero a costa de una mayor demanda computacional. Esta comparación subraya la importancia de seleccionar la estructura de datos ideal según el tamaño del conjunto de datos y los objetivos del modelo propuesto.

2.7. Aprendizaje automático y clasificación

De acuerdo con Palacios Sáez-Guinea (2022), el aprendizaje automático (o *machine learning* en inglés) es una rama de la inteligencia artificial caracterizada por el entrenamiento de una máquina con el propósito de que sea capaz de aprender sin instrucciones explícitas. Esto significa que los sistemas no se programan específicamente para cada caso, sino que cuentan con algoritmos capaces de generalizar comportamientos a partir de datos o ejemplos. Se trata de un proceso

de conocimiento que busca obtener resultados en un ambiente cambiante, de manera similar al pensamiento en los seres racionales. Esta disciplina se utiliza en una amplia variedad de aplicaciones comerciales y científicas, que abarcan desde la conducción no tripulada de vehículos y el reconocimiento facial, hasta herramientas directamente ligadas al procesamiento de señales de audio, tales como asistentes virtuales inteligentes, traducción automatizada de voz y sistemas de reconocimiento de voz comerciales (Sandoval Serrano et al., 2018).

El aprendizaje automático puede clasificarse en varias categorías; sin embargo, para los propósitos de este trabajo conviene describir dos:

- Aprendizaje supervisado.
- Aprendizaje no supervisado.

Por un lado, el aprendizaje no supervisado es un método en el que el modelo busca generar conocimiento adaptándose a las observaciones de los datos de entrada, sin conocimiento previo. Su objetivo principal es encontrar patrones comunes que permitan clasificar los datos en grupos con características similares. No obstante, la información obtenida bajo este enfoque debe ser interpretada posteriormente por un humano para que resulte útil. En cambio, el aprendizaje supervisado se utiliza en aquellos datos que cuentan con información específica acerca de los resultados deseados. Es decir, se emplean datos etiquetados para que la máquina conozca los resultados que debe obtener y de esta manera los compare con los obtenidos realmente, permitiendo corregir errores y mejorar el proceso de entrenamiento (Palacios Sáez-Guinea, 2022).

El presente trabajo utiliza un enfoque de aprendizaje supervisado, el cual se detallará en la siguiente sección.

2.7.1. Aprendizaje supervisado

El aprendizaje supervisado es un enfoque del aprendizaje automático que utiliza datos etiquetados para entrenar a un sistema en la predicción de resultados, imitando el proceso de aprendizaje bajo la guía de un instructor (Alnuaimi & Albaldawi, 2024). En este método, el modelo

analiza dicha información previa para descubrir los patrones que le permitan clasificar los datos en las categorías correspondientes. Por consiguiente, la eficiencia del algoritmo se calcula midiendo la diferencia entre las etiquetas que asigna el modelo y las etiquetas originales de los datos. Una vez concluido el entrenamiento, el sistema adquiere la capacidad de clasificar muestras nunca antes vistas. El proceso general de este tipo de aprendizaje se muestra en la Figura 2.15.



Figura 2.15. Diagrama de flujo del proceso del aprendizaje supervisado. Adaptado (traducido) de Syed y Lokhande (2024).

Shaveta (2023) menciona que el aprendizaje supervisado puede agruparse en dos categorías principales de algoritmos:

- **Clasificación:** se utiliza cuando la variable de salida es categórica, es decir, que pertenece a grupos discretos como sí/no, correcto/incorrecto. Algunos de los métodos más comunes incluyen los bosques aleatorios, los árboles de decisión y las máquinas de vectores de soporte (SVM).
- **Regresión:** se aplica cuando existe una relación continua entre las variables de entrada y salida, permitiendo predecir valores numéricos como el clima o las tendencias del mercado. Entre sus algoritmos destacan la regresión lineal, los árboles de regresión y la regresión bayesiana.

En este contexto, el reconocimiento de hablantes comúnmente se plantea como una tarea de clasificación supervisada, en la que el modelo debe aprender a asignar cada muestra de audio a la identidad correspondiente. Para resolverla, existen diferentes algoritmos en la literatura, entre los que destacan los bosques aleatorios, los k -vecinos más cercanos (k -NN) y las SVM. En este trabajo se seleccionaron principalmente las SVM debido a su efectividad al operar con espacios de características de dimensión media, como los coeficientes MFCC en su representación 1D. A continuación, se describe con mayor detalle el funcionamiento y relevancia de este clasificador.

2.7.2. Algoritmos de aprendizaje automático

Máquinas de vectores de soporte (SVM)

Las SVM son algoritmos de aprendizaje supervisado desarrollados en la década de 1990, cuyo principio fundamental es la generalización del clasificador de margen máximo. Aunque inicialmente este concepto requería que los datos fuesen linealmente separables, las versiones actuales permiten ajustar límites de clase no lineales, lo que expande su rango de aplicación. Estas máquinas están destinadas principalmente a tareas de clasificación binaria, es decir, donde existen dos clases, aunque su uso puede extenderse a problemas multiclase (Ramírez & Ramos, 2015). En la actualidad, este método se utiliza ampliamente en la clasificación de objetos, el reconocimiento facial y la categorización de texto, demostrando un rendimiento altamente eficiente en comparación con otros enfoques tradicionales (Zheng & Ding, 2018).

De acuerdo con Ramírez y Ramos (2015) y Zheng y Ding (2018), el funcionamiento de las SVM se fundamenta en los siguientes componentes esenciales:

- Hiperplano: es la frontera de decisión geométrica encargada de separar las distintas clases dentro del espacio de características.
- Vectores de soporte: son los puntos de datos pertenecientes a la muestra que se encuentran más cercanos al hiperplano, determinando de manera directa su posición y orientación óptima.

- **Margen:** consiste en la distancia máxima de separación perpendicular entre el hiperplano y los vectores de soporte más próximos de cada clase.
- **Funciones kernel:** son funciones matemáticas que permiten transformar datos originalmente no lineales a espacios de una dimensión mayor (espacio de características), lugar donde sí se vuelven linealmente separables.

El hiperplano óptimo se determina maximizando la distancia entre los puntos más cercanos de cada clase, conocidos como vectores de soporte. Este margen, definido como la distancia mínima entre la frontera y los ejemplos más próximos de cada grupo, debe ser lo más amplio posible. Por lo tanto, el hiperplano ideal es aquel que logra el máximo espacio de separación, garantizando una mejor capacidad de generalización del modelo. Véase un ejemplo en la Figura 2.16.

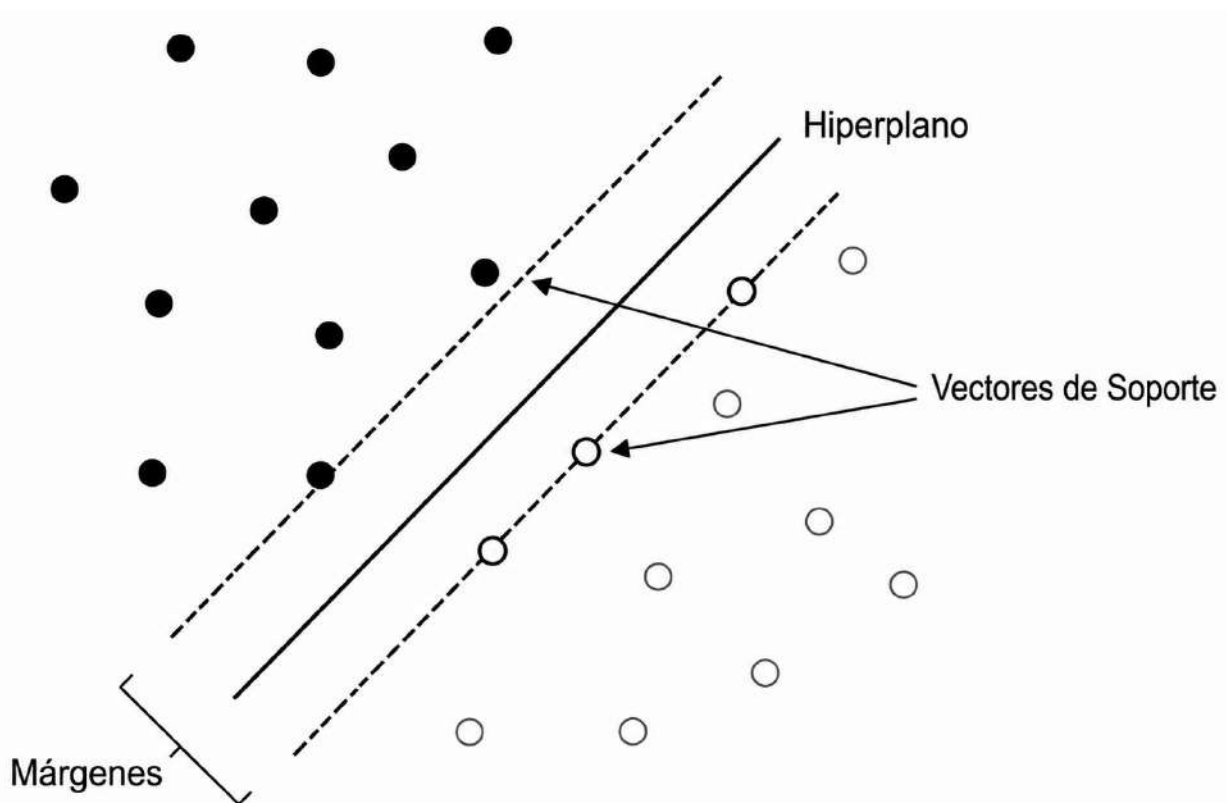


Figura 2.16. Representación de una máquina de vectores de soporte, mostrando el hiperplano, los vectores de soporte y los márgenes. Imagen adaptada (traducida) de Maia et al. (2020).

El proceso de construcción y entrenamiento de las SVM clásicas se divide metodológica-

mente en etapas secuenciales orientadas a maximizar la separabilidad de los datos. De acuerdo con la literatura técnica de Burges (1998), este flujo de trabajo consta de las siguientes fases fundamentales:

- Transformación de los datos (truco del kernel): se preparan los datos de entrada. Si no se pueden separar con una línea recta en su estado original, el algoritmo usa una función llamada kernel para proyectarlos a una dimensión mayor donde sí sea posible dividirlos de forma lineal.
- Búsqueda de la frontera (optimización): se calcula la línea o frontera de decisión ideal para dividir las clases de datos. El objetivo principal en esta etapa es lograr que dicha frontera tenga el mayor espacio libre o margen posible hacia ambos lados.
- Selección de puntos clave (vectores de soporte): el algoritmo identifica los datos más cercanos a la frontera de decisión. Estos puntos específicos son los únicos que sostienen y definen la posición final de la línea divisoria, mientras que el resto de las muestras se vuelven irrelevantes.
- Clasificación (fase de prueba): es la etapa operativa donde el modelo ya está entrenado. Cuando recibe un dato completamente nuevo, el sistema simplemente analiza de qué lado de la frontera geométrica se ubica y le asigna la etiqueta correspondiente. ““

A partir del desarrollo teórico y los experimentos presentados por Zheng y Ding (2018), se pueden fundamentar las siguientes ventajas operativas de esta técnica de aprendizaje supervisado:

- Eficiencia en espacios de alta dimensión: el algoritmo demuestra un rendimiento óptimo en tareas complejas como la categorización de texto, mostrando robustez al operar con datos de alta dimensionalidad.
- Maximización del margen: al centrarse en encontrar un hiperplano con el mayor margen de separación posible, el modelo ofrece una alta capacidad de generalización y reduce notablemente el riesgo de sobreajuste.

- Versatilidad mediante *kernels*: la implementación de funciones *kernel* permite al modelo adaptarse de forma efectiva tanto a problemas de separación lineales como no lineales.

A pesar de sus grandes ventajas, estos modelos también cuentan con algunas limitaciones. Por ejemplo, su costo computacional llega a ser elevado cuando existe una gran cantidad de ejemplos de entrenamiento. También se requiere una selección cuidadosa de hiperparámetros, como el parámetro C de penalización. Una elección inadecuada de estos valores puede comprometer la funcionalidad del algoritmo, afectando negativamente el rendimiento de la clasificación.

Aunque las SVM representan el núcleo de esta investigación por su efectividad en el manejo de vectores MFCC en 1D, no son la única alternativa dentro del aprendizaje supervisado. Para contrastar su desempeño y contar con puntos de referencia, también se consideraron otros algoritmos de clasificación, como bosques aleatorios y k -vecinos más cercanos (k -NN), los cuales, aunque más simples en su planteamiento, permiten evaluar la solidez y ventajas específicas de la SVM frente a diferentes enfoques. Estos se mencionan brevemente a continuación:

Bosques aleatorios

De acuerdo con Alnuaimi y Albaldawi (2024), el algoritmo de bosques aleatorios (*Random Forest* o RF) es un método de aprendizaje por conjunto y un algoritmo supervisado ampliamente utilizado en la ciencia de datos para resolver problemáticas de clasificación y regresión. Este enfoque constituye una extensión de los árboles de decisión tradicionales, en el cual se construyen múltiples árboles con los datos de entrenamiento para formar una estructura similar a un bosque. Ante la llegada de una nueva muestra, esta se procesa a través de los árboles generados, determinando la etiqueta final mediante un voto por mayoría en tareas de clasificación o a través del promedio en el caso de la regresión. Un ejemplo de esto puede apreciarse en la Figura 2.17. Por lo general, los bosques aleatorios ofrecen un rendimiento superior al de un único árbol de decisión, aunque su precisión puede verse influenciada por las características específicas de los datos de entrada.

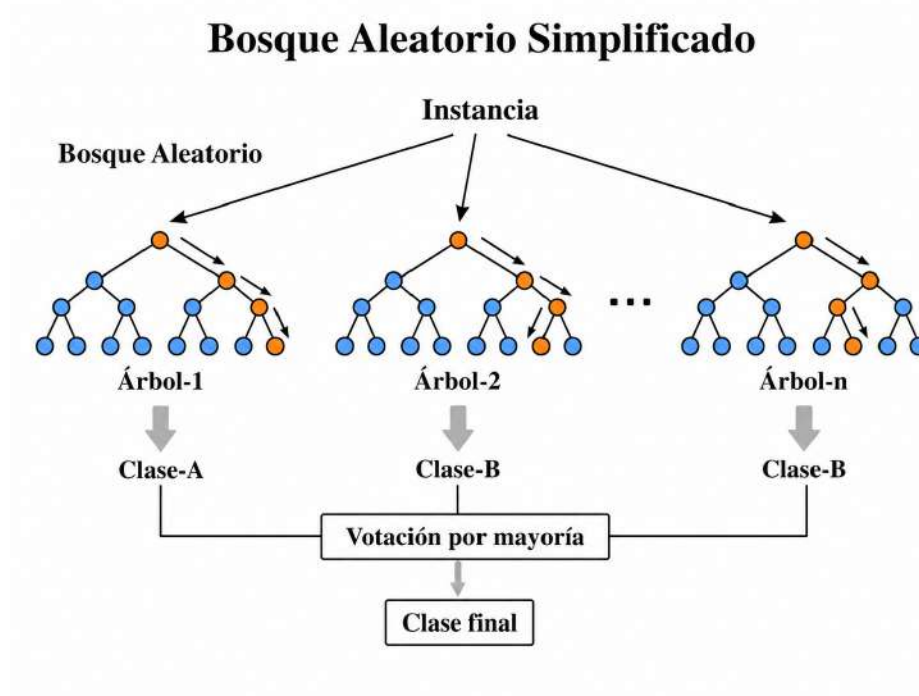


Figura 2.17. Representación del algoritmo de bosques aleatorios. La instancia es clasificada como clase A por un árbol y como clase B por dos árboles; por lo tanto, gracias a la mayoría de votos, será clasificada finalmente como clase B. Imagen adaptada (traducida) de Spotfire Community (s.f.).

Respecto a las características de este algoritmo, los autores destacan las siguientes ventajas:

- Capacidad con grandes volúmenes de datos: el algoritmo demuestra un desempeño óptimo al manejar conjuntos de datos extensos y complejos.
- Estabilidad estructural: el modelo exhibe una alta robustez ante la presencia de ruido ambiental y valores atípicos en las muestras de entrenamiento.
- Mitigación del sobreajuste: al promediar los resultados de múltiples estimadores individuales, se reduce notablemente el riesgo de que el sistema se memorice los datos.

Además, Alnuaimi y Albaldawi (2024) señalan las siguientes limitaciones y desventajas de este método:

- Prolongado período de entrenamiento: el proceso de ajuste y construcción de múltiples árboles requiere un tiempo significativamente mayor en comparación con otros algoritmos.

- Alta complejidad del modelo: la arquitectura interna exige un elevado consumo de recursos de memoria y costo computacional.
- Retraso en la fase operativa: debido a la necesidad de evaluar la muestra en todo el conjunto de árboles, las predicciones en tiempo real tienden a ser lentas.

A pesar de estas limitaciones, los bosques aleatorios cuentan con una amplia variedad de aplicaciones prácticas en escenarios reales, tales como el cálculo de niveles de riesgo en solicitudes de crédito bancario y la predicción de fallas mecánicas en los componentes de motores automotrices.

***k*-vecinos más cercanos (*k*-NN)**

Como también explican Alnuaimi y Albaldawi (2024), el algoritmo *k*-vecinos más cercanos (*k*-NN) es un método de aprendizaje automático que compara las métricas de similitud de los nuevos datos respecto a las muestras existentes para clasificarlas. Su funcionamiento se basa en una votación por mayoría simple entre los vecinos más próximos a cada punto de información (véase Figura 2.18). Uno de los principales desafíos en este algoritmo radica en la selección adecuada del número de vecinos con el que se tomará la decisión de clasificación. Debido a esto, el modelo *k*-NN es ampliamente utilizado en la estimación estadística y el reconocimiento de patrones, basándose en las relaciones de proximidad de los objetos.

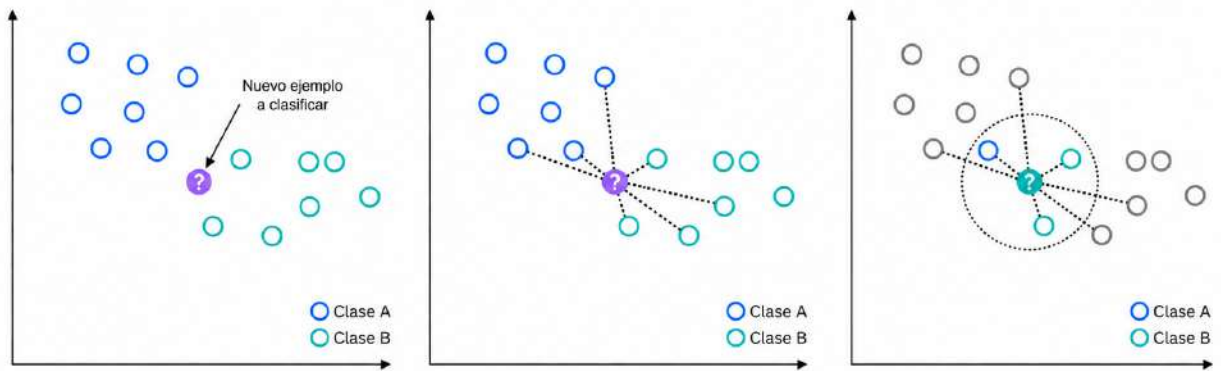


Figura 2.18. Representación del algoritmo (k -NN). Un nuevo elemento aparece y se clasifica en la clase B al definir $k = 3$. Imagen adaptada (traducida) de IBM (s.f.).

Las ventajas de este algoritmo, descritas por Alnuaimi y Albaldawi (2024), son las siguientes:

- Simplicidad y eficiencia: el modelo destaca por su facilidad de implementación y su rendimiento óptimo incluso en conjuntos de datos de gran escala.
- Bajo costo de construcción: la fase de preparación o construcción del modelo prácticamente no requiere inversión de recursos computacionales.
- Versatilidad adaptativa: el algoritmo demuestra una alta flexibilidad, siendo de gran utilidad tanto para tareas de regresión como de clasificación.

Sin embargo, los autores también indican algunas limitaciones que deben considerarse:

- Alta demanda computacional: a medida que el conjunto de entrenamiento incrementa su tamaño, el proceso se vuelve computacionalmente intensivo, disminuyendo su exactitud ante la presencia de ruido o características irrelevantes.
- Alto consumo de memoria: el sistema requiere almacenar la totalidad de los datos de entrenamiento en la memoria para poder operar.
- Retraso en la fase operativa: el proceso de predicción puede volverse lento dependiendo del volumen y la dimensión del conjunto de datos analizado.

A pesar de estas limitaciones, el algoritmo k -NN cuenta con diversas aplicaciones prácticas, entre las que destacan la búsqueda industrial de tareas similares, la detección de escritura a mano, el reconocimiento de imágenes y video, el análisis de inventario, los sistemas de recomendación y el diagnóstico médico.

2.8. Redes neuronales y aprendizaje profundo

2.8.1. Redes neuronales artificiales

López y Fernández (2008) explican que las redes neuronales artificiales (RNA) intentan emular el comportamiento del cerebro humano, el cual se caracteriza por el aprendizaje a través de la experiencia y la extracción de conocimiento genérico a partir de un conjunto de datos. En su obra se destaca que existen diferencias fundamentales entre el procesamiento de una computadora convencional y el cerebro biológico; mientras que las computadoras presentan habitualmente una arquitectura basada en un microprocesador muy rápido capaz de ejecutar instrucciones complejas en serie, el cerebro está compuesto por millones de procesadores elementales, o neuronas biológicas, interconectadas entre sí que operan bajo un esquema masivamente paralelo. Estas neuronas biológicas no requieren ser programadas, sino que aprenden a partir de los estímulos que reciben de su entorno.

Bajo este principio, y como los autores señalan, las RNA están constituidas por neuronas artificiales, las cuales representan dispositivos simples de cálculo o procesamiento que, a partir de un vector de entrada procedente del mundo exterior o de estímulos recibidos de otras neuronas, procesan la información para proporcionar una respuesta única o salida. A partir de esta estructura, López y Fernández (2008) caracterizan tres tipos fundamentales de unidades o capas dentro de la arquitectura de la red:

- Unidades de entrada: reciben señales directas desde el entorno, provenientes de sensores o archivos de almacenamiento que contienen los patrones de aprendizaje.

- Unidades de salida: envían su señal directamente fuera del sistema una vez que ha finalizado el tratamiento y procesamiento de la información.
- Unidades ocultas: reciben estímulos y emiten respuestas exclusivamente dentro del sistema, sin mantener contacto con el exterior, siendo las encargadas de llevar a cabo el procesamiento básico y establecer la representación interna de los datos.

A partir de esta distribución, se conforma una arquitectura capaz de modelar relaciones complejas, cuya estructura se ejemplifica en la Figura 2.19. En este modelo, el vector de entrada está constituido por un conjunto de valores numéricos denotados desde X_1 hasta X_n . Cada uno de estos valores es transmitido hacia los procesadores elementales o neuronas de la red, representados individualmente por los nodos $P_1, P_2, \dots, P_n$. Dentro de cada neurona, las entradas se combinan mediante pesos asignados y se someten a una función de activación para transferir el resultado hacia la capa siguiente, generando finalmente una salida estimada que se identifica con la variable Y_k . De este modo, las RNA constituyen la base del aprendizaje profundo y son ampliamente utilizadas gracias a su capacidad de modelar funciones no lineales complejas y procesar grandes volúmenes de información.

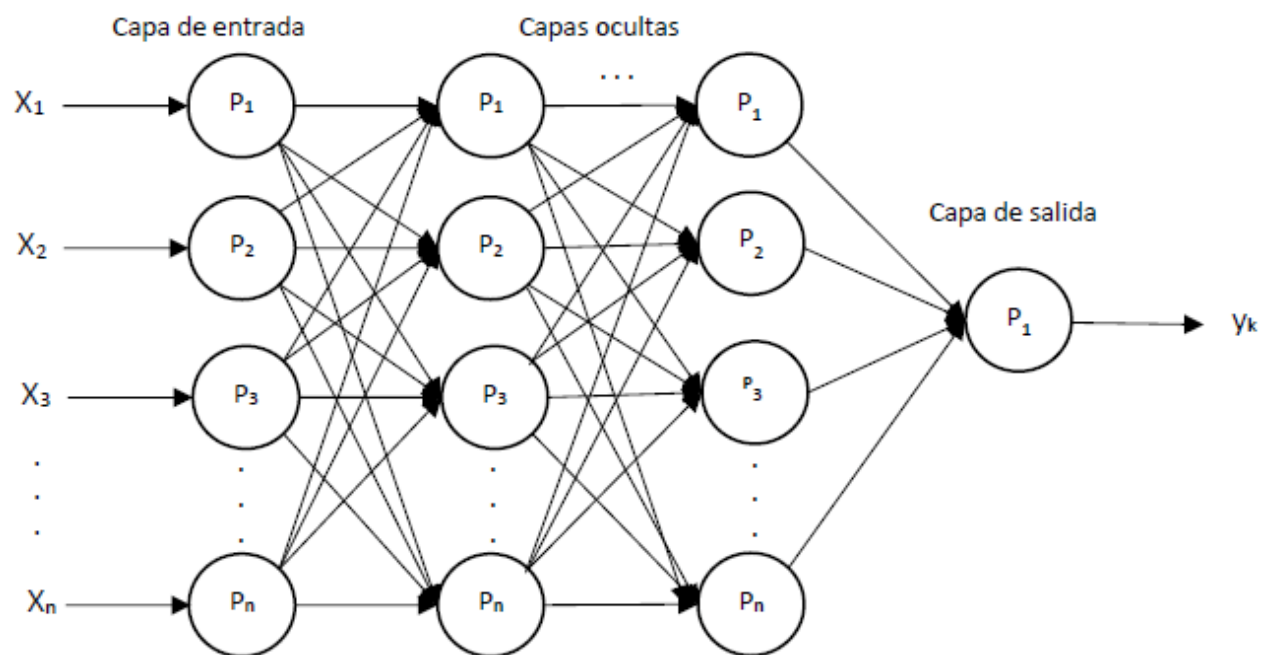


Figura 2.19. Estructura de una red neuronal artificial. Obtenido de Galindo et al. (2020).

2.8.2. Aprendizaje profundo

Los algoritmos de aprendizaje profundo utilizan estructuras de RNA y recursos computacionales para construir redes más grandes y complejas. El término “profundo” proviene del uso de múltiples capas ocultas, normalmente más de tres, orientadas al procesamiento de conjuntos de datos extensos. Esta disciplina constituye un área de la inteligencia artificial que busca emular el procesamiento de información del cerebro humano y su capacidad para la toma de decisiones mediante técnicas estadísticas y modelado predictivo. En este enfoque, los problemas de clasificación se abordan a través del entrenamiento de modelos que identifican las características distintivas de cada clase para aprender a partir de ellas. Posteriormente, el sistema predice la etiqueta para un nuevo objeto y su exactitud es evaluada con base en la tasa de aciertos sobre dichas etiquetas (Alnuaimi & Albaldawi, 2024).

Por su parte, Huang (2024) mencionan que el aprendizaje profundo ha avanzado notablemente en el campo del reconocimiento y la clasificación de imágenes mediante el desarrollo de arquitecturas multicapa sofisticadas. Estas redes son capaces de extraer información de alta di-

mensionalidad, permitiendo que la computadora asimile las relaciones internas de los datos para generar representaciones abstractas más ricas y complejas. En este contexto, las redes neuronales convolucionales (CNN, por sus siglas en inglés: *Convolutional Neural Networks*) representan una de las estructuras más utilizadas para tareas de procesamiento visual, las cuales serán abordadas detalladamente en la siguiente sección.

Redes neuronales convolucionales

Las redes neuronales convolucionales representan un avance clave dentro de las técnicas de aprendizaje profundo para abordar tareas de reconocimiento y clasificación de imágenes. Las arquitecturas de las CNN poseen la capacidad de extraer de manera autónoma características de las imágenes, abarcando desde rasgos simples hasta patrones complejos, mediante la construcción de múltiples capas especializadas, como la capa convolucional, la capa de agrupación y la capa completamente conectada. Gracias a este enfoque estructurado, el modelo evita una gran cantidad de trabajo manual dedicado a la extracción de características durante la fase de preprocesamiento de las imágenes (Huang, 2024).

Ahlawat et al. (2020) explican que una red neuronal convolucional (CNN) es, en esencia, una variante de la red de perceptrón multicapa (MLP, por sus siglas en inglés: *Multilayer Perceptron*). El procesamiento en una CNN se inspira en el sistema visual humano y en cómo las personas perciben e identifican objetos; por ejemplo, el entrenamiento visual mediante la observación repetida de imágenes permite generar predicciones precisas sobre elementos nunca antes vistos. Una CNN opera bajo este mismo principio y destaca por su capacidad para analizar imágenes visuales, logrando proveer una exactitud de reconocimiento considerable incluso con conjuntos de datos de entrenamiento pequeños.

De acuerdo con los autores, el uso de un modelo CNN ofrece ventajas metodológicas clave frente a los esquemas tradicionales de MLP:

- Integración de etapas: el modelo unifica los pasos de extracción de características y clasificación en un solo flujo de trabajo, requiriendo un esfuerzo mínimo en el preprocesamiento

de las imágenes.

- Extracción automática: las CNN extraen de forma automatizada características ricas e interrelacionadas a partir de las muestras de entrada, prescindiendo del diseño manual de rasgos.
- Invarianza espacial: el sistema es capaz de reconocer los patrones sin importar si el objeto en la imagen está rotado, inclinado o movido de su posición original, gracias a que la red aprovecha la estructura geométrica y la relación entre los píxeles.

Como detallan Ahlawat et al. (2020), una CNN básica incluye tres componentes esenciales: la capa de convolución, la capa de agrupación (*pooling*) y la capa de salida, siendo la capa de agrupación opcional en ciertos diseños. Una arquitectura típica orientada a la clasificación se compone de una capa de entrada, múltiples capas ocultas, basadas en la repetición secuencial de operaciones de convolución, normalización y agrupación, una capa completamente conectada y una capa de salida.

La interconexión parcial entre los nodos de las diferentes etapas permite procesar la información a través de pequeñas subregiones de la imagen de entrada denominadas campos receptivos (*receptive fields*), emulando la respuesta de un estímulo visual hacia la siguiente capa. La estructura típica y el flujo de procesamiento de esta arquitectura se ejemplifican en la Figura 2.20.

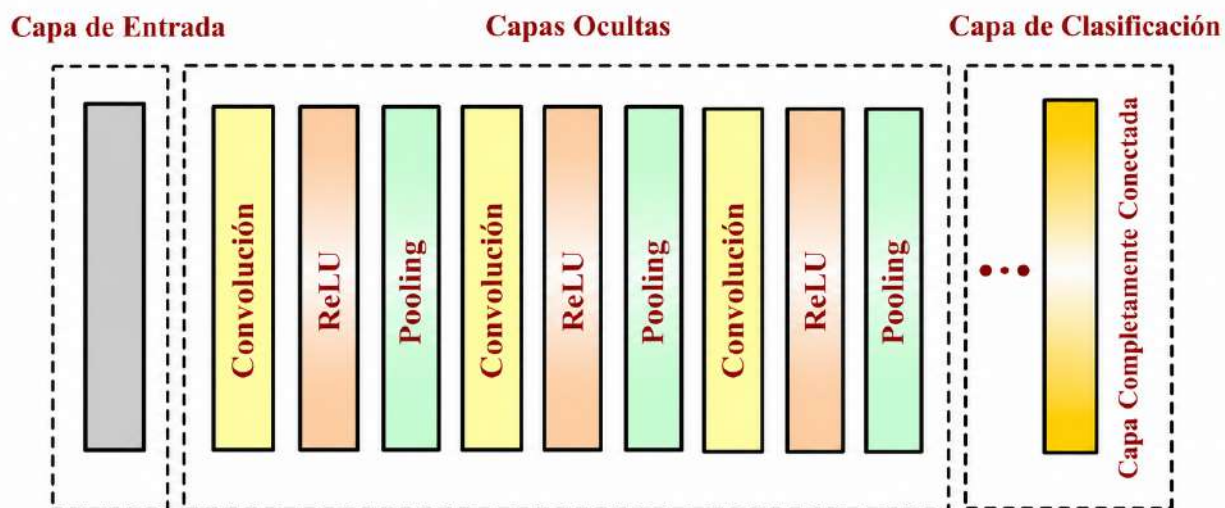


Figura 2.20. Estructura típica de una arquitectura de red neuronal convolucional (CNN). Imagen adaptada (traducida) de Ahlawat et al. (2020).

De acuerdo con el desglose presentado por Ahlawat et al. (2020), el funcionamiento de cada una de las capas que componen esta arquitectura se describe a continuación:

- **Capa de entrada:** es la encargada de cargar y almacenar los datos iniciales del sistema. En esta etapa se definen las dimensiones geométricas de la imagen de entrada, especificando su altura, anchura y el número de canales de color.
- **Capas ocultas:** actúan como la columna vertebral de la arquitectura, siendo las responsables de ejecutar el proceso global de extracción de características mediante la aplicación sucesiva de funciones de convolución, activación y agrupación.
- **Capa convolucional:** se posiciona como la primera capa operativa sobre la imagen de entrada con el objetivo de extraer sus rasgos esenciales. Su computación matemática se inspira en la corteza visual de los animales, procesando la información visual a partir de pequeñas subregiones del campo de entrada.
- **Capa de agrupación (*pooling*):** se integra estratégicamente entre las capas convolucionales para reducir el tamaño de las imágenes y hacer el procesamiento más eficiente. Para ello, el

modelo divide la información en pequeñas regiones aisladas, como una cuadrícula, y extrae un único valor representativo de cada una, lo que permite conservar las características más importantes, descartar los datos redundantes y ayudar a controlar el sobreajuste.

- Capa de activación: al igual que en las redes neuronales convencionales, se incluye con el propósito fundamental de introducir no linealidad en el sistema, permitiendo el modelado de relaciones complejas.
- Capa de clasificación: constituye la última etapa de la arquitectura y se compone de una red completamente conectada (*fully connected*). Esta capa se encarga de calcular las probabilidades de las clases deseadas, combinando la totalidad de las características abstractas que la red aprendió en los niveles previos.

En esta investigación, las características acústicas MFCC se emplean como imágenes bidimensionales que representan la evolución temporal de la voz, donde cada archivo de audio equivale a una muestra visual única. Al transformar estas matrices en imágenes, es posible aprovechar las capacidades de las redes convolucionales para identificar patrones espaciales y temporales en dichas representaciones. Con este propósito, se seleccionó la arquitectura VGG16, una red convolucional cuya configuración estructural se ilustra en la Figura 2.21. Este modelo se utilizará como base para llevar a cabo el entrenamiento y la clasificación de las imágenes correspondientes a los MFCC de los distintos individuos incluidos en el presente trabajo.

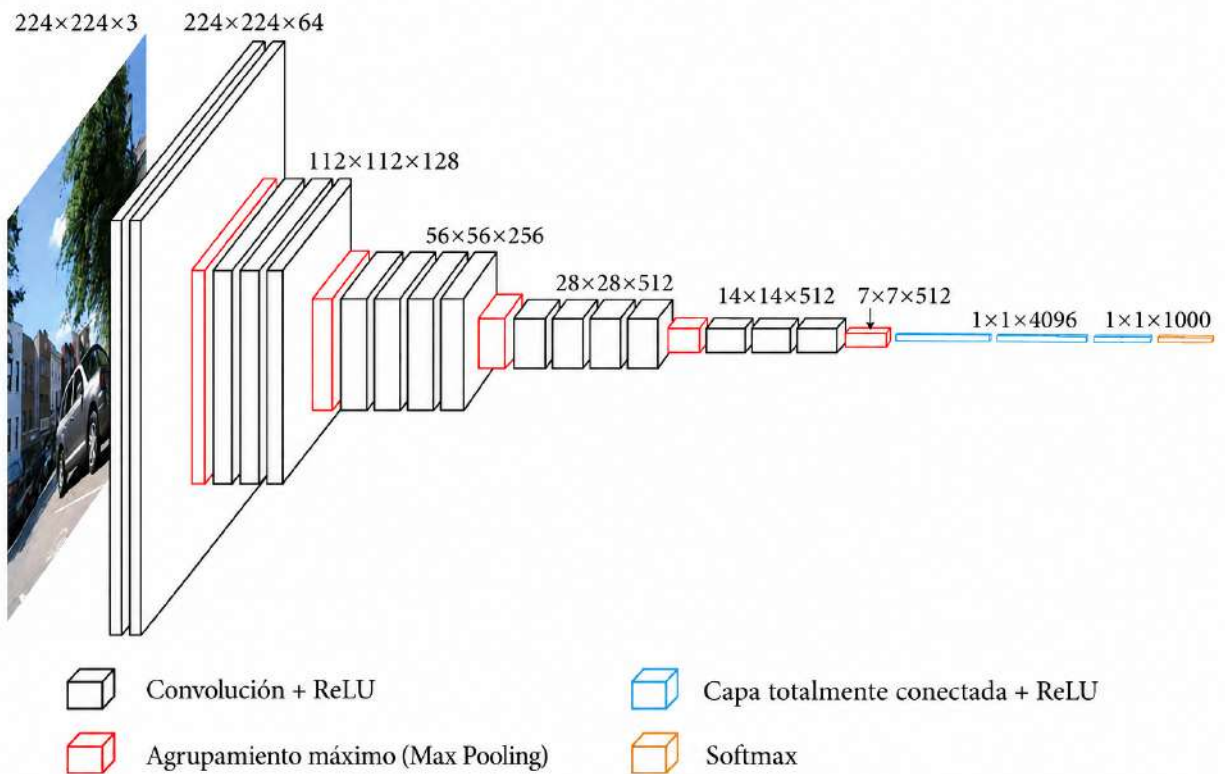


Figura 2.21. Estructura interna y dimensiones de la arquitectura VGG16, detallando la distribución de sus 16 capas con pesos constituidas por bloques de convolución, activación de unidad lineal rectificadora (ReLU, por sus siglas en inglés: *Rectified Linear Unit*) y capas completamente conectadas. Imagen adaptada (traducida) de Zhu y Han (2022).

Capítulo 3

Estado del arte

3.1. Enfoques de clasificación supervisada para reconocer al hablante

3.1.1. Enfoques basados en representaciones en formato 1D

Tradicionalmente, el análisis de los coeficientes cepstrales en la escala de Mel (MFCC, por sus siglas en inglés: *Mel Frequency Cepstral Coefficients*) se ha abordado mediante un enfoque clásico estructurado en forma de vectores. Aunque la extracción de estos descriptores genera originalmente una matriz bidimensional que relaciona los atributos de la voz con el tiempo, los métodos convencionales requieren transformar dicha estructura para poder procesarla. Por lo general, esto se logra promediando los coeficientes a lo largo del tiempo o aplanando la matriz para convertirla en una única fila de datos por cada archivo de audio.

Bajo este esquema, cada muestra de audio se representa como un vector plano dentro de un arreglo de propiedades acústicas. Estos vectores se introducen directamente como entradas en clasificadores estadísticos tradicionales, tales como las máquinas de vectores de soporte (SVM, por sus siglas en inglés: *Support Vector Machines*), el algoritmo de k -vecinos más cercanos (k -NN, por sus siglas en inglés: *k-Nearest Neighbors*) o redes neuronales de baja profundidad. La popularidad de dicha estrategia radica principalmente en su bajo costo computacional y en su alta eficacia cuando se trabaja en entornos controlados o con conjuntos de datos reducidos.

Ortega-Mendoza et al. (2007) proponen un método para incluir cambios de tiempo dentro de los MFCC, con la idea de observar cómo evoluciona la señal de voz. En este estudio se menciona que los coeficientes comunes no logran capturar por sí solos cómo varía la emisión de una persona a

lo largo del tiempo, por lo que hace falta agregar datos dinámicos que ayuden al sistema a distinguir mejor a cada hablante.

La metodología se desarrolló con el conjunto de datos DIMEX100 y consistió en calcular cuatro medidas estadísticas básicas para cada coeficiente y sus derivadas temporales: el valor mínimo, el valor máximo, el promedio y la varianza (véase Tabla 3.1).

Tabla 3.1. Medidas estadísticas aplicadas a los cambios de los coeficientes. Adaptado (traducido) de Ortega-Mendoza et al. (2007).

Descripción	Cálculo	Número de características
Valor máximo de los cambios del coeficiente	$\hat{\Delta}c_i = \max_{k=2}^N (c_{ik} - c_{i(k-1)})$	16
Valor mínimo de los cambios del coeficiente	$\check{\Delta}c_i = \min_{k=2}^N (c_{ik} - c_{i(k-1)})$	16
Valor promedio de los cambios del coeficiente	$\tilde{\Delta}c_i = \frac{1}{N-1} \sum_{k=2}^N (c_{ik} - c_{i(k-1)})$	16
Varianza de los cambios del coeficiente	$\Delta_v c_i = \frac{1}{N-1} \sum_{k=2}^N (c_{ik} - c_{i(k-1)} - \tilde{\Delta}c_i)^2$	16

El desempeño de estos atributos se analizó mediante los clasificadores de Bayes ingenuo (*Naïve Bayes*) y máquinas de vectores de soporte. Durante las pruebas se modificó el tamaño del contexto temporal tomado en cuenta y se evaluaron combinaciones con menos datos usando solo las estadísticas calculadas. El sistema alcanzó una exactitud máxima de 97.56 % al utilizar el modelo SVM con la configuración de cuatro estadísticas sobre dieciséis coeficientes.

Este trabajo señala que el análisis de las variaciones temporales en los coeficientes ayuda directamente a reconocer mejor a los hablantes. Metodológicamente, este antecedente sirve como fundamento conceptual para la presente investigación, debido a que expone cómo se deben diseñar y extraer manualmente las estadísticas para resumir el tiempo en un vector plano, volviendo viable el uso de modelos tradicionales como la SVM, que no procesan matrices de forma nativa.

En cambio, Chowdhury y Ross (2020) presentan una forma de reconocer hablantes enfocada en mejorar la precisión cuando los audios tienen ruido de fondo o distorsiones. Para resolver este problema, se propone unir dos tipos de rasgos de la voz que se complementan: los MFCC, que representan cómo escucha el oído humano el sonido, y los coeficientes de predicción lineal (LPC, por sus siglas en inglés: Linear Predictive Coding), que describen la forma física en que el tracto vocal produce el habla.

El núcleo de este trabajo está en su estrategia de procesamiento. Los autores explican que los datos de audio no se pueden tratar como si fueran imágenes comunes, ya que el tiempo y las frecuencias cambian de formas distintas. Por ello, en lugar de mezclar estos datos en un análisis de dos dimensiones, el modelo procesa los atributos manteniendo un formato secuencial de una sola dimensión (1D) marco por marco. Para lograrlo, se utiliza una red neuronal convolucional diseñada para operar en una sola dimensión (*1D-CNN*), la cual revisa los coeficientes acústicos a lo largo del tiempo sin cambiar su estructura original. El flujo general del modelo se puede ver en la Figura 3.1.

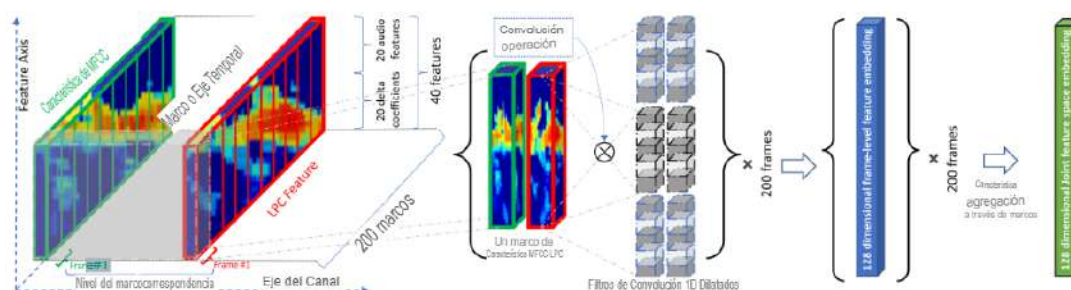


Figura 3.1. Flujo general del modelo de fusión de características mediante una red 1D-Triplet-CNN. Imagen adaptada (traducida) de Chowdhury y Ross (2020).

Los experimentos demostraron que este enfoque basado en rasgos de una sola dimensión supera por mucho a los modelos que intentan introducir los datos de audio directamente en redes de imágenes de dos dimensiones (2D), ya que estas últimas suelen confundir al sistema y causar errores en el aprendizaje.

De este modo, el trabajo de Chowdhury y Ross (2020) respalda la utilidad y la vigencia de los MFCC organizados en secuencias unidimensionales, validando la relevancia de mantener la

estructura lineal en esta investigación. Su éxito demuestra que organizar la señal en vectores marco por marco es una estrategia sólida y eficiente para asegurar la estabilidad del clasificador, incluso ante factores adversos como el ruido y la distorsión del entorno.

Por otro lado, Liu et al. (2021) proponen una metodología para mejorar la precisión en la verificación de hablantes modificando la manera en que se calculan los rasgos de la voz. La idea central es conservar la estructura y la facilidad de lectura de los MFCC comunes, pero permitiendo que el módulo encargado de extraer estos componentes tenga la flexibilidad de aprender y ajustarse automáticamente a los datos en lugar de operar con parámetros fijos predefinidos desde el inicio. En este estudio se demuestra que perfeccionar la etapa donde se captura la información acústica es tan importante para el rendimiento general como el diseño del clasificador final.

Para dicho fin, el sistema toma los pasos normales del cálculo de los MFCC, tales como el enventanado, la transformada de Fourier, el banco de filtros Mel y la transformada discreta del coseno, y los convierte en bloques adaptables dentro de la red neuronal. Al inicio del experimento, estos componentes usan los valores fijos tradicionales, pero a medida que el sistema entrena, se van afinando solos mediante el aprendizaje de la propia red. De esta forma, la computadora aprende a cambiar los filtros matemáticos para que se adapten de mejor manera a las propiedades reales de la voz de los usuarios.

Los autores reportaron mejoras relativas de entre el 6.7 % y el 9.7 % en la reducción de errores al evaluar el modelo con los conjuntos de datos VoxCeleb1 (audios de famosos extraídos de internet) y SITW (Speakers in the Wild; grabaciones tomadas de medios en condiciones del mundo real). Esta mejora en el rendimiento, la cual se detalla a lo largo del entrenamiento en la Figura 3.2, se alcanzó manteniendo estrictamente los datos organizados en un formato secuencial de una sola dimensión (1D). Esto demuestra que es posible hacer más preciso el sistema optimizando los rasgos matemáticos a lo largo del tiempo sin necesidad de alterar la dimensión original del audio.

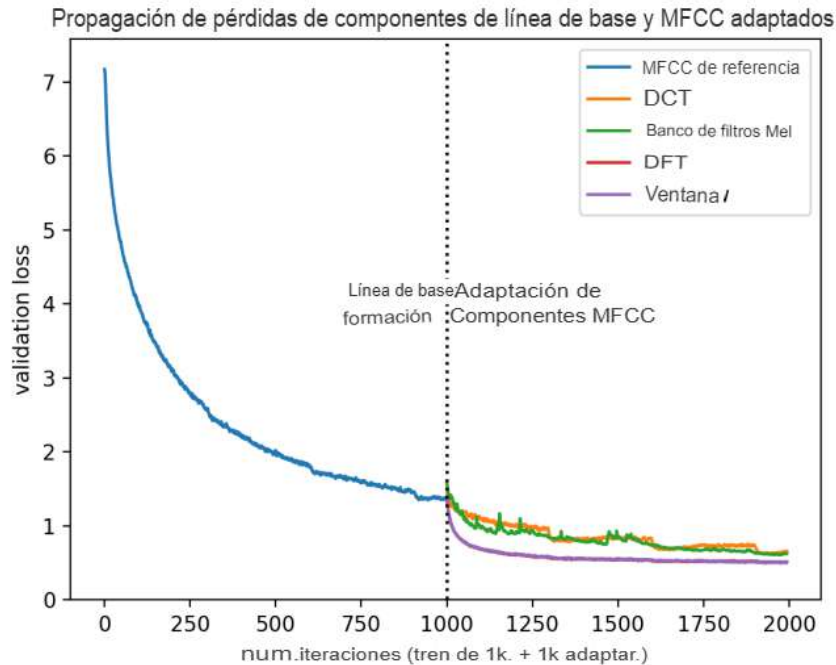


Figura 3.2. Gráfica del rendimiento de los datos con método tradicional y parámetros aprendibles. Imagen adaptada (traducida) de Liu et al. (2021).

Así, el trabajo de Liu et al. (2021) manifiesta el alcance de las representaciones en formato 1D, sirviendo como una justificación metodológica clave para dar soporte a la contraparte clásica de esta tesis. Su éxito experimental demuestra que los vectores de MFCC comunes, cuando se procesan mediante un esquema optimizado en una sola dimensión, constituyen una base altamente competitiva. Asimismo, este antecedente proporciona una referencia de alto rendimiento en el ámbito unidimensional, evidenciando que la permanencia de la estructura secuencial a través del tiempo resulta en una estrategia eficiente para distinguir la identidad de los sujetos.

Por último, Zhang y Liu (2023) proponen una metodología para mejorar la precisión en el reconocimiento de hablantes mediante la combinación de diferentes tipos de características de la voz. Este planteamiento surge debido a que un solo descriptor de forma aislada no siempre logra cubrir toda la información relevante del habla, por lo cual se busca demostrar que la integración inteligente de varios rasgos acústicos permite construir un sistema más robusto y con una mayor capacidad para adaptarse a nuevas voces.

Para cumplir este objetivo, el modelo toma diferentes representaciones de la voz en un for-

mato de una sola dimensión (1D) a lo largo del tiempo, incluyendo los MFCC tradicionales, bancos de filtros (FBANK), coeficientes de predicción lineal (LPCC) y espectrogramas convencionales. A nivel de marco temporal, estas señales se procesan incorporando unidades compartidas denominadas *shared units*, las cuales permiten que el sistema aprenda y aproveche la información común entre las distintas ramas de audio. Posteriormente, la secuencia de cuadros se condensa mediante un resumen estadístico basado en promedios y variaciones que se apoya en un mecanismo de atención para resaltar las partes más importantes del audio. El resultado de este proceso es un *x-vector*, el cual consiste en un vector de características de tamaño fijo que representa de manera compacta la identidad del hablante. El flujo general de este procesamiento se ilustra en la Figura 3.3.

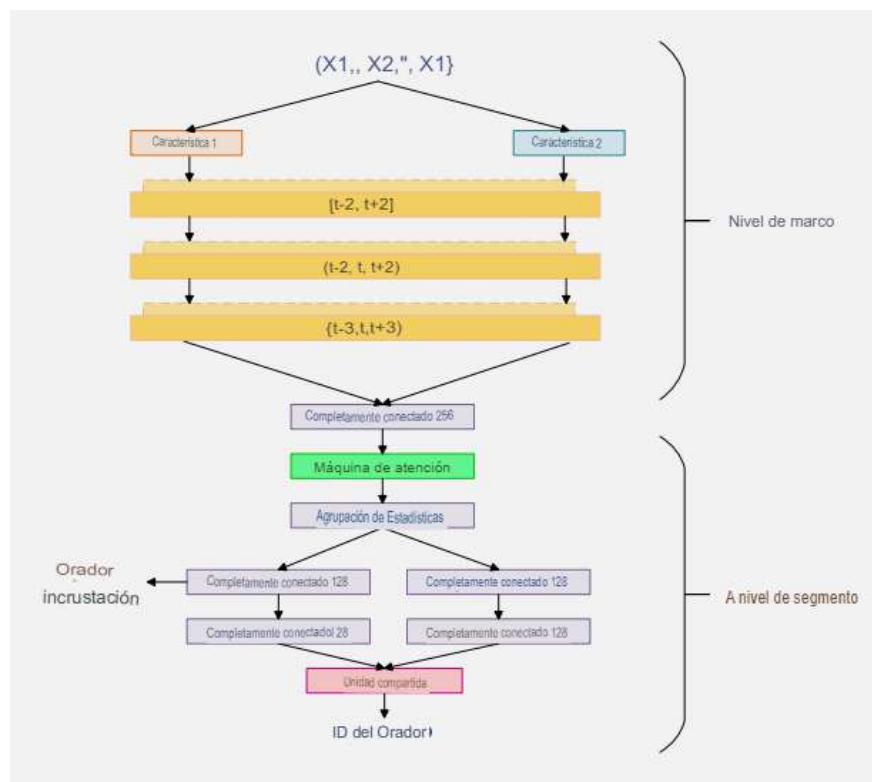


Figura 3.3. Sistema de reconocimiento del hablante propuesto por Zhang y Liu (2023), adaptado (traducido) de la obra original.

En los resultados, las pruebas experimentales demostraron que combinar los atributos reduce significativamente los errores en comparación con los modelos que usan un solo tipo de rasgo. Al evaluar el sistema con el conjunto de datos VoxCeleb1, el modelo alcanzó una tasa de error de

0.98 % utilizando los *x-vectors* generados. Además, al aplicar una técnica de fusión de puntajes finales de salida llamada *score fusion*, los fallos disminuyeron todavía más hasta registrar un error de 0.65 %, confirmando que la arquitectura multitarea aporta ventajas frente a enfoques individuales.

El trabajo de Zhang y Liu (2023) respalda la viabilidad de las representaciones en formato 1D, aportando un marco de comparación de vanguardia para la presente investigación. Su éxito demuestra que los vectores de MFCC comunes, al complementarse con otros rasgos dentro de una sola dimensión, constituyen una base sólida y de alta precisión. Aunque el enfoque del presente desarrollo no se centra en combinar múltiples descriptores, sino en comparar de forma aislada los MFCC unidimensionales frente a su versión bidimensional, esta obra respalda la idea de que explorar alternativas en la representación de los datos es un camino totalmente válido para avanzar en el reconocimiento de hablantes.

3.1.2. Enfoques basados en representaciones en formato 2D

Además del uso común de los MFCC en formato lineal, existe la posibilidad de tratarlos como imágenes mediante una representación en dos dimensiones (2D). La idea consiste en organizar los coeficientes como matrices numéricas que sirven de entrada para redes neuronales convolucionales (CNN, por sus siglas en inglés: *Convolutional Neural Networks*). Este cambio de perspectiva permite aprovechar la capacidad nativa de las CNN para detectar patrones espaciales y texturas complejas, aplicando de manera directa sobre datos acústicos aquellas técnicas de aprendizaje profundo que originalmente se desarrollaron para el análisis visual.

Wirdiani et al. (2024) presentan un sistema de reconocimiento de hablantes que apunta a preservar la estructura bidimensional de tiempo y frecuencia de los MFCC en lugar de convertirlos en vectores planos. La motivación de los autores radica en que aplanar la estructura implica la pérdida de patrones espectro-temporales que son importantes para distinguir a cada hablante. De este modo, lo que buscan demostrar es que usar las matrices de MFCC en dos dimensiones (2D), combinadas con arquitecturas convolucionales, puede mejorar la exactitud frente a los métodos tradicionales que usan la representación vectorial.

El desarrollo del modelo involucra una red neuronal convolucional residual (*Residual Convolutional Neural Network*, ResCNN) diseñada para procesar las características acústicas organizadas en tensores numéricos bidimensionales con un tamaño fijo de 160×64 (véase Figura 3.4). Durante el entrenamiento, el sistema congela las primeras capas de la red y añade componentes de ajuste fino que se optimizan mediante la técnica de minería de tripletas en línea (*online triplet mining*). Esta estrategia selecciona conjuntos de tres ejemplos compuestos por una muestra de referencia, una positiva del mismo hablante y una negativa de otro individuo, forzando al modelo a separar los atributos de personas distintas y agrupar las muestras del mismo sujeto para generar un vector de incrustación (*embedding*) de 128 dimensiones. Finalmente, estas representaciones numéricas se clasifican en memoria mediante el algoritmo de k -vecinos más cercanos (k -NN). En los experimentos, se usaron los datos del conjunto TIMIT y grabaciones en vivo tomadas con teléfonos inteligentes para validar el sistema.

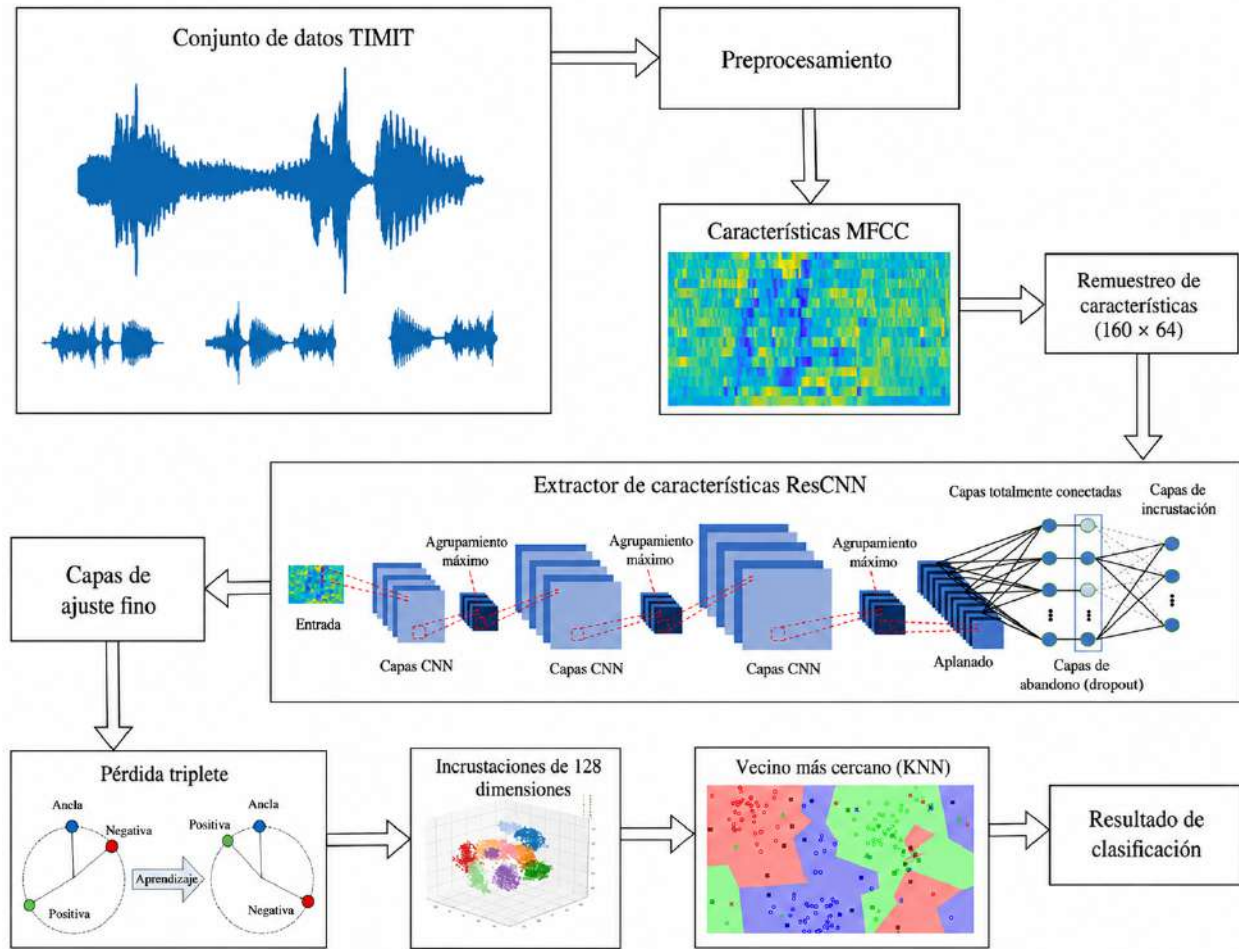


Figura 3.4. Sistema de reconocimiento del hablante propuesto por Wirdiani et al. (2024), adaptado (traducido) de la obra original.

Los resultados reportados demostraron que evaluar el modelo con 25 épocas de entrenamiento reduce significativamente los fallos. El sistema obtuvo un 94.66 % de exactitud al utilizar las muestras del conjunto TIMIT, y alcanzó un 96.00 % de exactitud en las pruebas con grabaciones en vivo, superando en ambos escenarios a los modelos comparativos que emplean los datos en formato aplanado. Este resultado apoya la premisa de que mantener la organización matricial de los MFCC permite capturar de mejor manera las propiedades acústicas que caracterizan a cada hablante.

Para el presente desarrollo, este estudio tiene una importancia especial porque evidencia que la organización en 2D de los MFCC ofrece ventajas claras sobre la representación vectorial tradicional. Sin embargo, existe una diferencia metodológica fundamental entre ambos enfoques.

Mientras que Wirdiani et al. (2024) procesan las muestras estrictamente como arreglos numéricos en la memoria de la computadora, en esta investigación se da un paso hacia la visión por computadora real al transformar físicamente las matrices de MFCC en archivos de imagen reales con mapas de color definidos. Esta diferencia permite justificar la incorporación de la arquitectura VGG16, diseñada originalmente para el análisis visual nativo. De este modo, en este trabajo se propone un marco experimental controlado para evaluar si transformar los MFCC en imágenes reales y clasificarlas mediante visión por computadora ofrece un mejor rendimiento frente al uso de los vectores tradicionales en una sola dimensión (1D).

Por su parte, Youngsik y Bang (2021) abordaron el problema del reconocimiento de emociones en el habla mediante un enfoque basado en MFCC en formato bidimensional. El interés de los autores se centró en explorar cómo estas representaciones permiten capturar matices acústicos asociados a variaciones emocionales en la voz, tales como cambios sutiles en la entonación y el timbre. Este objetivo se planteó debido a que las transiciones emocionales se manifiestan con mayor claridad en las variaciones de energía distribuidas en el espectro frecuencial a lo largo del tiempo, lo que exige el uso de descriptores capaces de capturar los cambios de la voz en el tiempo y la frecuencia de forma conjunta.

El sistema se implementó mediante una red neuronal convolucional bidimensional (2D-CNN) estándar, la cual fue alimentada directamente con las propiedades acústicas en matrices de dos dimensiones (2D). En lugar de utilizar arquitecturas híbridas o excesivamente complejas, la metodología se centró en dejar que el modelo aprendiera los patrones espectro-temporales de la voz directamente de la matriz, tratándolos conceptualmente como si fueran las texturas de una imagen. Para estandarizar las entradas del clasificador, se aplicó un preprocesamiento que consistió en recortar los audios extensos a una duración fija de tres segundos y rellenar con ceros (*zero-padding*) los segmentos más cortos. El flujo general de este procesamiento se ilustra en la Figura 3.5.

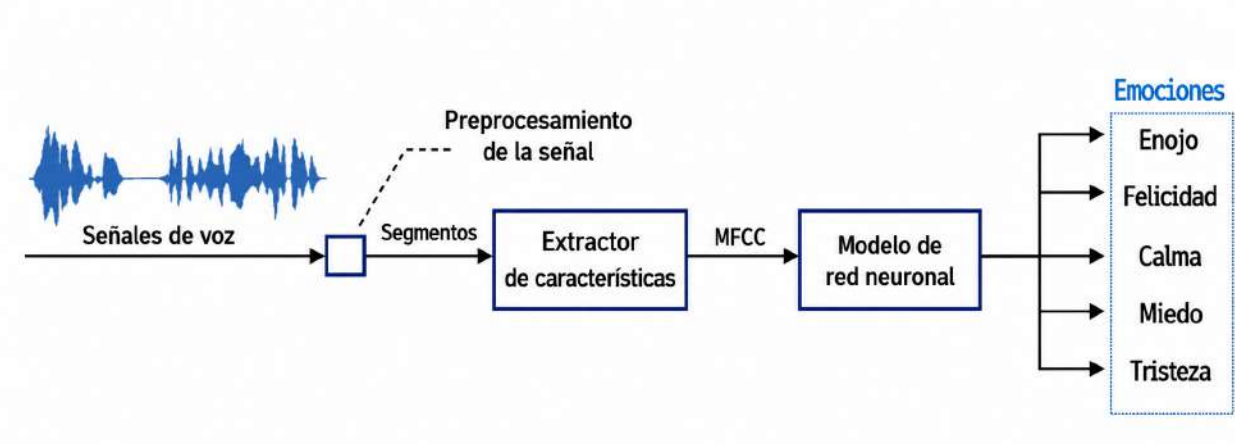


Figura 3.5. Sistema de reconocimiento del hablante propuesto por Youngsik y Bang (2021), adaptado (traducido) de la obra original.

Las pruebas realizadas sobre el conjunto de datos RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) arrojaron un promedio de 88.54 % de exactitud al clasificarse cinco estados afectivos diferenciados por género. Este desempeño demostró que mantener la estructura bidimensional de los rasgos acústicos, sin recurrir a un aplanamiento de los datos, otorga una alta capacidad para identificar señales complejas en la voz. Asimismo, el análisis determinó que añadir capas secuenciales adicionales o procesos de reducción excesivos terminaba por degradar el rendimiento debido a la pérdida de información temporal, consolidando el uso directo de redes convolucionales puras como una alternativa altamente competitiva.

Este estudio aporta al presente desarrollo un antecedente de alto valor metodológico al demostrar que los patrones de energía de los MFCC pueden explotarse con éxito mediante el análisis espacial en dos dimensiones. Al igual que en el trabajo de Youngsik y Bang (2021), la presente investigación aprovecha la estructura de los MFCC para estudiar cómo se distribuyen sus propiedades a lo largo del tiempo, compartiendo incluso los mismos criterios de preparación de datos al estandarizar los audios a una duración fija de tres segundos mediante el recorte de segmentos largos y la aplicación de técnicas de relleno para los fragmentos cortos. Sin embargo, existe una diferencia metodológica fundamental entre ambos proyectos: mientras que los autores procesan las características estrictamente como arreglos numéricos en la memoria de la computadora, en

este trabajo se propone una transición hacia la visión por computadora al transformar físicamente esas matrices en archivos de imagen reales con mapas de color definidos. Esta diferencia justifica la incorporación de la arquitectura VGG16, diseñada para el análisis visual, estableciendo un marco experimental controlado para evaluar si el flujo de clasificación de imágenes optimiza la discriminación de características frente al enfoque unidimensional (1D).

Finalmente, Costantini et al. (2023) exploraron el reconocimiento de hablantes desde una perspectiva con base en representaciones visuales de la señal de voz. El objetivo de su investigación fue analizar si el uso de distintas formas de representación bidimensional exportadas como archivos de imagen físicos, en particular los espectrogramas y los gráficos de MFCC (*MFCC-graphs*), podía aprovechar las capacidades de las redes neuronales convolucionales. Este enfoque parte de la hipótesis de que transformar el audio en un artefacto visual real permite aplicar de manera directa técnicas de visión por computadora para identificar los rasgos característicos que distinguen la identidad de cada sujeto.

El desarrollo metodológico consistió en evaluar el rendimiento de ambos tipos de imágenes utilizando la arquitectura preentrenada AlexNet y una red personalizada de cinco capas convolucionales denominada CNN1 (véase Figura 3.6). Para llevar a cabo los experimentos, se utilizó el conjunto de datos DEMoS, el cual comprende 8869 segmentos de audio pertenecientes a 58 hablantes expresados en diferentes estados emocionales, lo que permitió validar los modelos bajo un escenario de alta variabilidad en el timbre y la entonación de la voz.

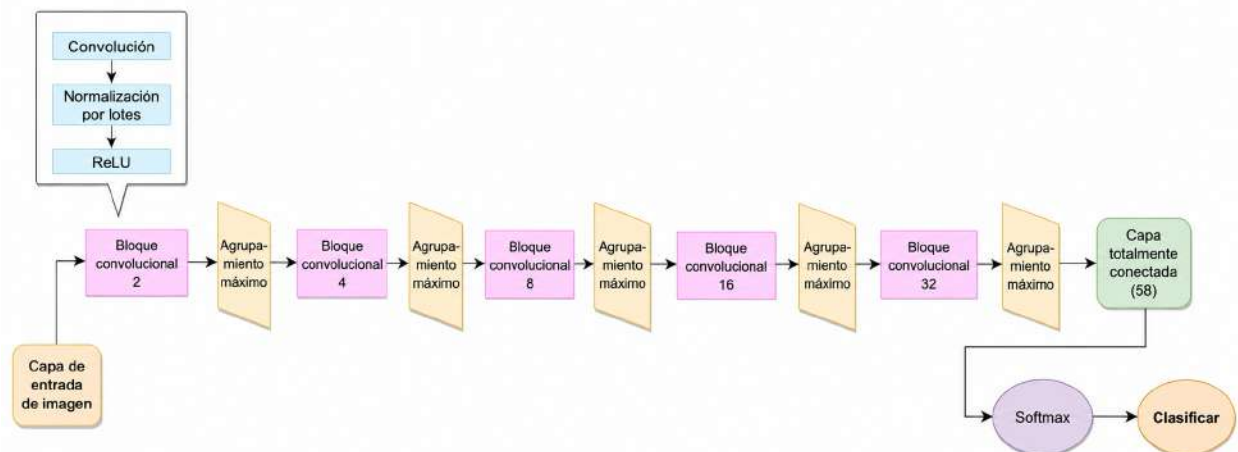


Figura 3.6. Flujo de procesamiento de la red convolucional personalizada CNN1. Imagen adaptada (traducida) de Costantini et al. (2023).

Durante el procesamiento, los autores generaron y guardaron archivos físicos en formato PNG (Portable Network Graphics), evaluando de forma sistemática tanto variantes en escala de grises como a color para cada una de las representaciones visuales generadas (véase Figura 3.7).

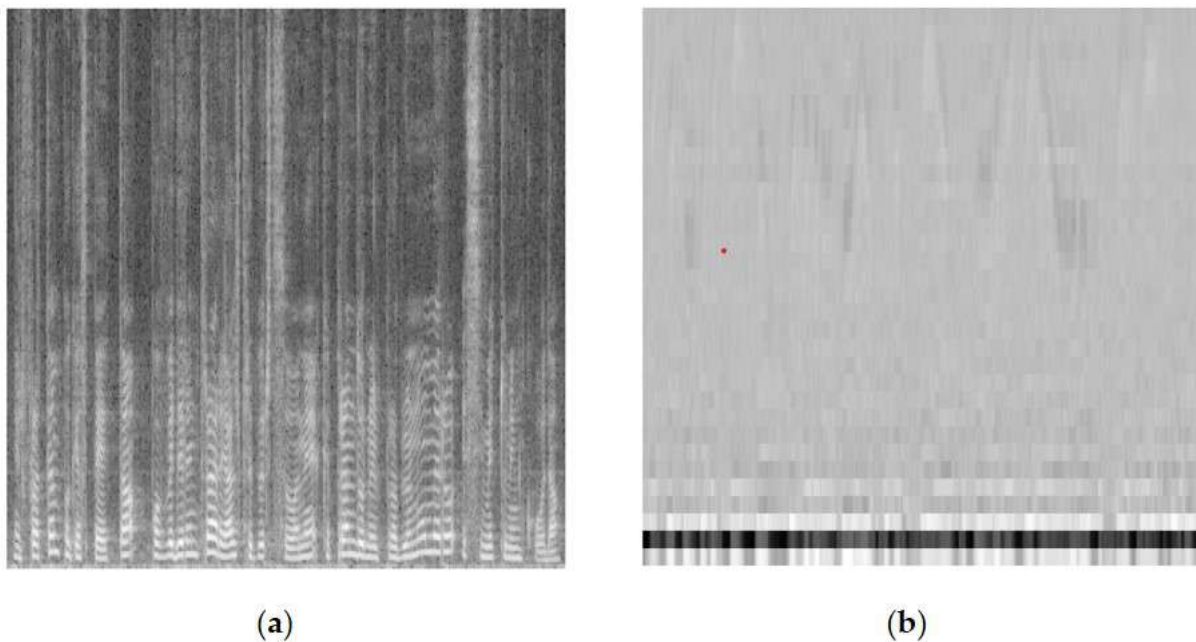


Figura 3.7. Ejemplos de las representaciones visuales empleadas en el trabajo de Costantini et al. (2023): (a) espectrograma en escala de grises y (b) gráfico de MFCC del mismo segmento de audio.

Sus resultados mostraron que el uso de espectrogramas en escala de grises ofreció el mejor

desempeño general. La red personalizada CNN1 alcanzó una exactitud del 90.15 % con espectrogramas en escala de grises y un 88.85 % en su versión a color, mientras que para los gráficos de MFCC obtuvo un 83.17 % a color y un 82.27 % en escala de grises. De manera similar, la arquitectura AlexNet registró un 89.28 % de exactitud con la representación de espectrogramas en escala de grises y un 83.43 % con las imágenes a color de los MFCC. Estas cifras confirman que, aunque ambas representaciones visuales en sus distintas modalidades cromáticas son aptas para la tarea de reconocimiento, el formato y la distribución de la información gráfica influyen de forma directa en la capacidad de generalización del clasificador.

El trabajo mencionado es especialmente relevante para la presente investigación debido a que comparte el mismo principio metodológico de transformar las matrices de MFCC en archivos de imagen reales para someterlas a un flujo de visión por computadora. Sin embargo, existe una brecha metodológica importante: Costantini et al. (2023) limitan su análisis a la comparación del rendimiento entre formatos de imagen y alternativas cromáticas bidimensionales, pero no establecen un contraste frente a métodos clásicos de procesamiento de señales. Por lo tanto, en este estudio se propone abordar este tema mediante el desarrollo de un marco experimental controlado, cuyo propósito es evaluar de forma directa si la transición hacia un flujo de clasificación visual utilizando la arquitectura VGG16 ofrece ventajas reales en la exactitud frente al uso tradicional en una sola dimensión (1D).

3.1.3. Comparación de representaciones acústicas multimodales

Martínez Mascorro y Aguilar Torres (2013) propusieron un sistema de reconocimiento de hablantes diseñado para mitigar el impacto de las variaciones en la emisión de la voz mediante el uso de un esquema multimodal. El objetivo del estudio se centró en explorar cómo la integración de tres algoritmos de parametrización distintos permite capturar propiedades acústicas complementarias para robustecer el proceso de identificación. Para cumplir este propósito, el desarrollo metodológico involucró el análisis de los MFCC en formato unidimensional (1D), los parámetros cepstrales basados en sub-bandas (*Sub-Band Cepstral Coefficients*) y las representaciones visua-

les obtenidas por medio de espectrogramas convencionales de dos dimensiones (2D). Las señales de audio se procesaron de forma paralela utilizando un enfoque híbrido de clasificación, donde los vectores de coeficientes alimentaron a un modelo de máquinas de vectores de soporte (SVM), mientras que los espectrogramas se analizaron mediante la aplicación de filtros de Gabor y un clasificador basado en la distancia euclidiana, lo que facilitó la producción de un conjunto independiente de candidatos en cada rama antes de que se emitiera un veredicto en conjunto.

Para evaluar el rendimiento de dicha propuesta, los experimentos se llevaron a cabo empleando una base de datos compuesta por grabaciones de 19 personas, donde cada participante registró muestras bajo diferentes tonos de voz que simulaban cambios voluntarios e involuntarios en el habla, tales como voz natural, voz modificada intencionalmente y voz ronca causada por enfermedad. Los resultados demostraron que cada tipo de descriptor aporta ventajas específicas al sistema, evidenciando que los MFCC vectoriales destacan por su bajo costo computacional y su estabilidad, mientras que los espectrogramas ofrecen una representación más rica en el dominio conjunto del tiempo y la frecuencia, lo que optimiza la discriminación de las identidades bajo condiciones de voz alterada. A partir de estas observaciones, se concluyó que la combinación de características de diferente índole disminuye los errores de clasificación causados por las variaciones físicas en el tracto vocal de los usuarios.

Este trabajo posee una relevancia fundamental para la presente investigación debido a que integra el análisis de representaciones unidimensionales y bidimensionales dentro de un mismo entorno experimental. No obstante, la propuesta de Martínez Mascorro y Aguilar Torres (2013) se orienta estrictamente hacia un esquema de fusión multimodal, donde se busca combinar la información de diferentes descriptores para compensar sus limitaciones individuales y obtener una decisión final basada en el conjunto de las votaciones. En contraste, en este estudio se propone un enfoque puramente comparativo y controlado que busca aislar las variables mediante el uso exclusivo de los MFCC, revisando de manera directa el comportamiento de su estructura 1D frente a su transformación en archivos de imagen reales analizados mediante la arquitectura VGG16, midiendo de forma objetiva la efectividad de cada alternativa.

3.2. Conjuntos de datos

Para el desarrollo de esta investigación se recurrió a conjuntos de datos de voz de acceso público, ampliamente utilizados en el ámbito académico para tareas de reconocimiento automático de hablantes. La elección de estos conjuntos de datos responde a la necesidad de trabajar con información útil y que permitiese comparar resultados en distintos escenarios de prueba.

3.2.1. Conjunto de datos DIMEX100

El primer conjunto de datos seleccionado corresponde a DIMEX100, el cual se compone de grabaciones de hablantes nativos del español mexicano y fue desarrollado por la Universidad Nacional Autónoma de México, cuyo acceso y documentación se encuentran disponibles en su repositorio oficial¹. Este conjunto de datos se caracteriza por contener señales de audio capturadas en condiciones de laboratorio altamente controladas, utilizando micrófonos de alta fidelidad y manteniendo un nivel de ruido ambiental mínimo (Pineda et al., 2004). Estas propiedades lo convierten en un punto de referencia para evaluar el desempeño de los modelos en un entorno limpio y uniforme.

3.2.2. Conjunto de datos VoxCeleb1

El segundo conjunto de datos empleado es VoxCeleb, una colección a gran escala obtenida de forma automatizada a partir de videos de la plataforma YouTube, diseñada específicamente para tareas de reconocimiento y verificación de identidades vocales². En la literatura de esta plataforma se identifican dos versiones principales, donde la primera de ellas, denominada VoxCeleb1, incluye clips de voz pertenecientes a 1251 celebridades distintas, mientras que la versión subsecuente, conocida como VoxCeleb2, amplía la cobertura a más de 6000 hablantes (Nagrani et al., 2017). Para los fines de este estudio se seleccionó la primera versión debido a que abarca una alta va-

¹Enlace de descarga de DIMEX100: <https://turing.iimas.unam.mx/luis/DIME/CORPUS-DIMEX.html>

²Enlace de descarga de VoxCeleb: <https://www.robots.ox.ac.uk/vgg/data/voxceleb/>

riabilidad en acentos, expresiones espontáneas, ruido de fondo y condiciones de grabación reales, lo que contrasta de manera directa con la naturaleza controlada del primer conjunto de datos en español mexicano.

3.2.3. Estadísticas de los conjuntos de datos

Con el fin de asegurar la uniformidad en el procesamiento, las señales de voz se estandarizaron en sus valores. En la Tabla 3.2 se presentan las especificaciones de cada conjunto de datos, detallándose tanto el total de usuarios disponibles como el grupo seleccionado para las pruebas, así como la cantidad de archivos asignados por individuo junto con el tiempo de duración que registran en promedio las grabaciones.

Tabla 3.2. Resumen de los conjuntos de datos utilizados

Conjunto de datos	Personas totales	Personas consideradas	Audios por persona	Duración promedio
DIMEX100	100	51	50	9.59 s
VoxCeleb1	1211	51	50	8.16 s

Se decidió seleccionar únicamente a un grupo de 51 personas de cada base de datos. Esta elección se basó en dos motivos fundamentales, donde el primero responde a las limitaciones técnicas del equipo de cómputo disponible, lo que permite procesar una cantidad menor de archivos de forma más rápida. El segundo motivo obedece a un criterio de igualdad experimental, ya que al tener exactamente la misma cantidad de hablantes y de audios en ambos conjuntos de datos se logra una comparación justa entre las pruebas limpias en español y los entornos con ruido en inglés.

3.3. Discusión

Al analizar los trabajos de la literatura, se nota un camino claro que va desde el uso tradicional de los audios en una sola dimensión hasta el procesamiento de imágenes en dos dimen-

siones. En el formato unidimensional (1D), los estudios iniciales de Ortega-Mendoza et al. (2007) confirman que los coeficientes cepstrales en la escala de Mel (MFCC) son descriptores ligeros y eficientes. En dicha investigación se demuestra que estos rasgos estáticos pueden mejorar su rendimiento si se les añaden métricas estadísticas manuales, como el máximo, el mínimo, el promedio y la varianza, lo que ayuda a mantener la información sobre los cambios de la voz a lo largo del tiempo. Sin embargo, trabajar en una sola dimensión suele traer problemas de pérdida de precisión cuando las señales de audio se exponen a entornos reales con ruido de fondo o distorsiones en el canal.

Para resolver estas debilidades sin abandonar el formato lineal, los investigadores han diseñado diferentes estrategias de compensación que aumentan la complejidad del sistema. Por ejemplo, en el trabajo de Chowdhury y Ross (2020) se optó por juntar los MFCC con coeficientes de predicción lineal (LPC) mediante el uso de una red convolucional 1D entrenada con minería de tripletas, buscando que la combinación de características proteja la identidad del hablante en grabaciones degradadas. De manera diferente, Liu et al. (2021) plantean modificar la extracción misma de los MFCC al introducir filtros matemáticos entrenables dentro de la propia red neuronal, logrando que los coeficientes se adapten de forma automática al tipo de voz. Por su parte, Zhang y Liu (2023) desarrollaron un sistema multitarea avanzado que agrupa descriptores acústicos como MFCC, bancos de filtros, espectrogramas y coeficientes LPCC mediante mecanismos de atención para empaquetar toda la información en un vector de tamaño fijo. Aunque todos estos métodos ofrecen resultados competitivos, se observa que los enfoques basados en 1D necesitan procesos matemáticos o la combinación de varias variables para evitar que los datos importantes se pierdan al aplanar las matrices originales.

Como alternativa a estas dificultades, otros autores sugieren el uso del formato bidimensional (2D) para aprovechar de mejor manera el potencial de las redes convolucionales profundas. En este grupo metodológico, las investigaciones de Youngsik y Bang (2021) y de Wirdiani et al. (2024) demuestran que mantener los MFCC organizados en forma de matriz evita que se destruyan las relaciones naturales que existen entre el tiempo y la frecuencia de la señal de voz. A pesar de

esta ventaja espacial, en ambos proyectos los datos se procesan estrictamente como arreglos de números. Esto provoca que las características se traten de forma interna, por lo que no se aprovechan las funciones avanzadas de las arquitecturas de visión por computadora que están diseñadas para trabajar con archivos de imagen reales.

Es en este punto donde el estudio de Costantini et al. (2023) propone un cambio relevante al generar y guardar físicamente archivos de imagen reales en formato PNG, lo que permite evaluar el impacto visual de los mapas de color y las escalas de grises mediante el uso del modelo AlexNet. Sin embargo, este análisis enfocado en imágenes se limita a comparar formatos visuales entre sí, dejando de lado un contraste directo contra las representaciones en una sola dimensión.

De igual forma, el esquema desarrollado por Martínez Mascorro y Aguilar Torres (2013) destaca como el único antecedente que junta ambos mundos al procesar vectores e imágenes de forma paralela mediante comités de votación para disminuir los errores de clasificación, pero su diseño se enfoca en la fusión de algoritmos y no realiza una revisión aislada para explorar el valor individual de cada formato.

Para revisar de manera clara las características de las investigaciones analizadas, se presenta un resumen comparativo en la Tabla 3.3, donde se muestran las representaciones, los modelos y los resultados principales.

Tabla 3.3. Comparativa de los trabajos abarcados en el estado del arte relacionados con representaciones acústicas para reconocimiento de hablantes

Trabajo	Representación	Algoritmo / Modelo	Conjunto de datos	Contexto del trabajo	Métrica	Resultados principales
Ortega-Mendoza et al. (2007)	MFCC 1D e información dinámica	SVM y Bayes Ingenuo	DIMEX100	Extracción manual de máximo, mínimo, promedio y varianza para resumir el tiempo	Exactitud	97.56 %
Chowdhury y Ross (2020)	MFCC 1D y coeficientes LPC	<i>1D-Triplet-CNN</i>	Grabaciones con degradación	Unión de vectores marco por marco para dar robustez frente al ruido lineal	EER	EER del 12.00 %
Martínez Mascorro y Aguilar Torres (2013)	MFCC 1D, SBC y espectrogramas 2D	SVM y comités de fusión	Muestra propia con variaciones de voz	Combinación paralela de rasgos vectoriales y visuales mediante reglas de votación	Exactitud	Fusión unificada del 91.00 %
Liu et al. (2021)	MFCC 1D con filtros aprendibles	Arquitectura profunda DNN	VoxCeleb1 y SITW	Inclusión de parámetros entrenables en el enventanado para optimizar el descriptor	EER	Reducción de error del 9.70 %
Youngsik y Bang (2021)	MFCC 2D matricial numérico	Red Convolutional <i>2D-CNN</i>	Base de datos emocional	Análisis de matrices MFCC fijas a tres segundos tratadas como texturas numéricas	Exactitud	88.54 %
Zhang y Liu (2023)	MFCC 1D, Fbank, espectrogramas y LPCC	Extractor <i>x-vector</i> multi-tarea	VoxCeleb1	Agrupación de descriptores lineales mediante capas compartidas y mecanismos de atención	EER	EER del 0.65 % en fusión
Costantini et al. (2023)	Gráficos MFCC y espectrogramas 2D	Modelos CNN y AlexNet	Base de datos DEMoS	Generación y almacenamiento físico de archivos PNG para evaluar mapas de color	Exactitud	Espectrogramas visuales con 90.15 %
Wirdiani et al. (2024)	MFCC 2D matricial numérico	Modelo <i>Res-CNN</i> con <i>triplet mining</i>	TIMIT señales vivo y en	Conservación de la matriz tiempo-frecuencia en memoria caché aplicando minería de tripletas	Exactitud	96.00 %

Se puede comprobar que los trabajos revisados se concentran en la combinación de múltiples tipos de descriptores acústicos o en el uso de matrices numéricas internas, pero no realizan una comparación entre el formato lineal original de los MFCC y su transformación física en imágenes reales. Los proyectos que eligen el formato bidimensional suelen depender de procesos dentro de la memoria del sistema, dejando de lado las ventajas de usar archivos gráficos independientes que puedan someterse al análisis de redes profundas.

A grandes rasgos, el estado del arte actual no solo contextualiza los avances previos, sino que también fundamenta la pertinencia del estudio al demostrar la falta de un contraste directo entre ambos formatos. De este modo, se establece el vínculo con la metodología que se desarrolla en el siguiente capítulo, donde se plantea un entorno de pruebas controlado para medir de forma equitativa el desempeño de las alternativas 1D y 2D bajo las mismas condiciones experimentales.

Capítulo 4

Metodología

En este capítulo se describe la metodología utilizada para el desarrollo de la investigación, la cual se basa en el paradigma KDD (*Knowledge Discovery in Databases*) adaptado al contexto del procesamiento de señales de voz. La metodología KDD se refiere al proceso organizado para descubrir conocimiento a partir de repositorios de datos, permitiendo la identificación de patrones válidos, útiles y comprensibles (Díaz, 2024).

Como sostienen Maimon y Rokach (2010), este enfoque constituye un análisis exploratorio automatizado y un modelado de grandes conjuntos de datos desordenados. Para el presente trabajo, el flujo metodológico se organiza en dos grandes etapas transversales que abarcan la extracción de características acústicas y el diseño del entorno de clasificación (Figura 4.1).

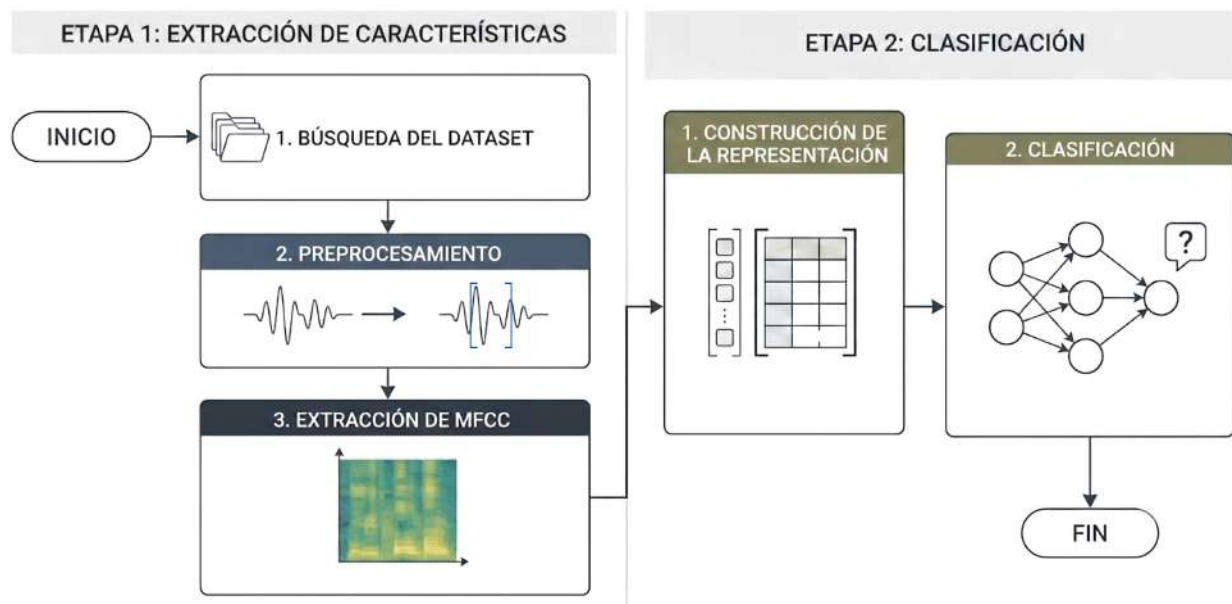


Figura 4.1. Metodología general dividida en dos etapas: extracción de características y clasificación.

4.1. Etapa 1: Extracción de características

4.1.1. Búsqueda y selección de los conjuntos de datos

La primera fase de la metodología consiste en la selección y el análisis de los conjuntos de datos que sirven como base para el entrenamiento y la evaluación de los modelos. Con el propósito de comparar el desempeño de los clasificadores bajo condiciones distintas de grabación, ruido de fondo y estilos de habla, se integraron dos corpus de audio de naturaleza complementaria conocidos como *DIMEX100* y *VoxCeleb*. El primer conjunto proporciona grabaciones de voz en español mexicano bajo un entorno controlado de estudio, mientras que el segundo grupo aporta registros en inglés obtenidos de videos de internet en escenarios no controlados, lo cual permite someter a los algoritmos a retos reales de variabilidad acústica y verificar la viabilidad de transferencia del sistema.

4.1.2. Preprocesamiento avanzado de la señal de audio

Esta fase consiste en preparar los archivos de audio crudos de manera que puedan ser procesados en condiciones iguales. El preprocesamiento inicia con un remuestreo digital para estandarizar la tasa de muestreo a un valor uniforme de 44,100 Hz en todos los registros. Posteriormente, se ejecuta una normalización de amplitud basada en el pico máximo para compensar las diferencias de ganancia de los micrófonos y se aplica un filtro de preénfasis para balancear el espectro de frecuencias altas de la voz.

Una vez preprocesada la señal, se realiza un recorte temporal exacto de tres segundos tomando como pivote el centro de la duración total del archivo, distribuyéndose 1.5 segundos a la izquierda y 1.5 segundos a la derecha de dicho eje simétrico. Para aquellos registros excepcionales que no alcanzan la duración mínima requerida, se implementa un algoritmo de relleno con ceros (*zero-padding*) en los espacios faltantes para asegurar vectores de longitud constante. Finalmente, la señal continua se segmenta en intervalos cortos mediante un proceso de inventanado fundamen-

tado en una ventana de Hamming, configurándose una longitud de transformada rápida de Fourier de 30 milisegundos y un tamaño de salto adaptado de forma equivalente a 30 milisegundos.

4.1.3. Extracción de MFCC

Una vez preprocesados los audios, se procede con la extracción de características. En esta etapa se aplica la técnica de los coeficientes cepstrales en la escala de Mel (MFCC, por sus siglas en inglés: *Mel Frequency Cepstral Coefficients*). Su relevancia radica en que permiten representar la señal de manera compacta, imitando la percepción humana de las frecuencias. Esta técnica es ampliamente utilizada en el procesamiento para el reconocimiento del habla y del hablante, como se ha analizado en capítulos anteriores.

La implementación de este proceso se realiza mediante el lenguaje de programación Python y diversas bibliotecas especializadas para el procesamiento de datos y señales. En particular, se utiliza la biblioteca *librosa* para la extracción de los coeficientes MFCC y sus deltas, *os* para la gestión de directorios y automatización del procesamiento de los archivos de audio, *NumPy* para la manipulación y transformación de arreglos numéricos que posteriormente se integran en matrices de características, *pandas* para la construcción de estructuras de datos tipo *DataFrame*, y *matplotlib.pyplot* para la generación de representaciones visuales bidimensionales a partir de las matrices de MFCC.

El resultado de este proceso es una matriz de MFCC por cada audio, donde las filas representan los distintos *frames* analizados y las columnas contienen los coeficientes numéricos obtenidos. Esta matriz constituye una representación intermedia de la voz que conserva información tanto espectral como temporal, y que sirve como base para la construcción de las representaciones unidimensional y bidimensional descritas en las siguientes secciones.

4.2. Etapa 2: Clasificación

En la segunda etapa se construyen las representaciones de los datos y se aplican los modelos de clasificación. Dentro de esta etapa se consideran dos enfoques distintos descritos mediante flujos de procesamiento específicos: la representación unidimensional de los MFCC, orientada al uso de clasificadores tradicionales como SVM, KNN y Bosques Aleatorios cuya secuencia metodológica se detalla en la Figura 4.2, y la representación bidimensional, donde los coeficientes se organizan como mapas tiempo–frecuencia para su procesamiento mediante redes neuronales convolucionales como VGG16, tal como se ilustra en el diagrama de la Figura 4.3. Esta estructura permite establecer un marco comparativo claro entre ambos enfoques bajo condiciones experimentalmente equivalentes, desglosándose las fases de la etapa dos en las siguientes subsecciones.

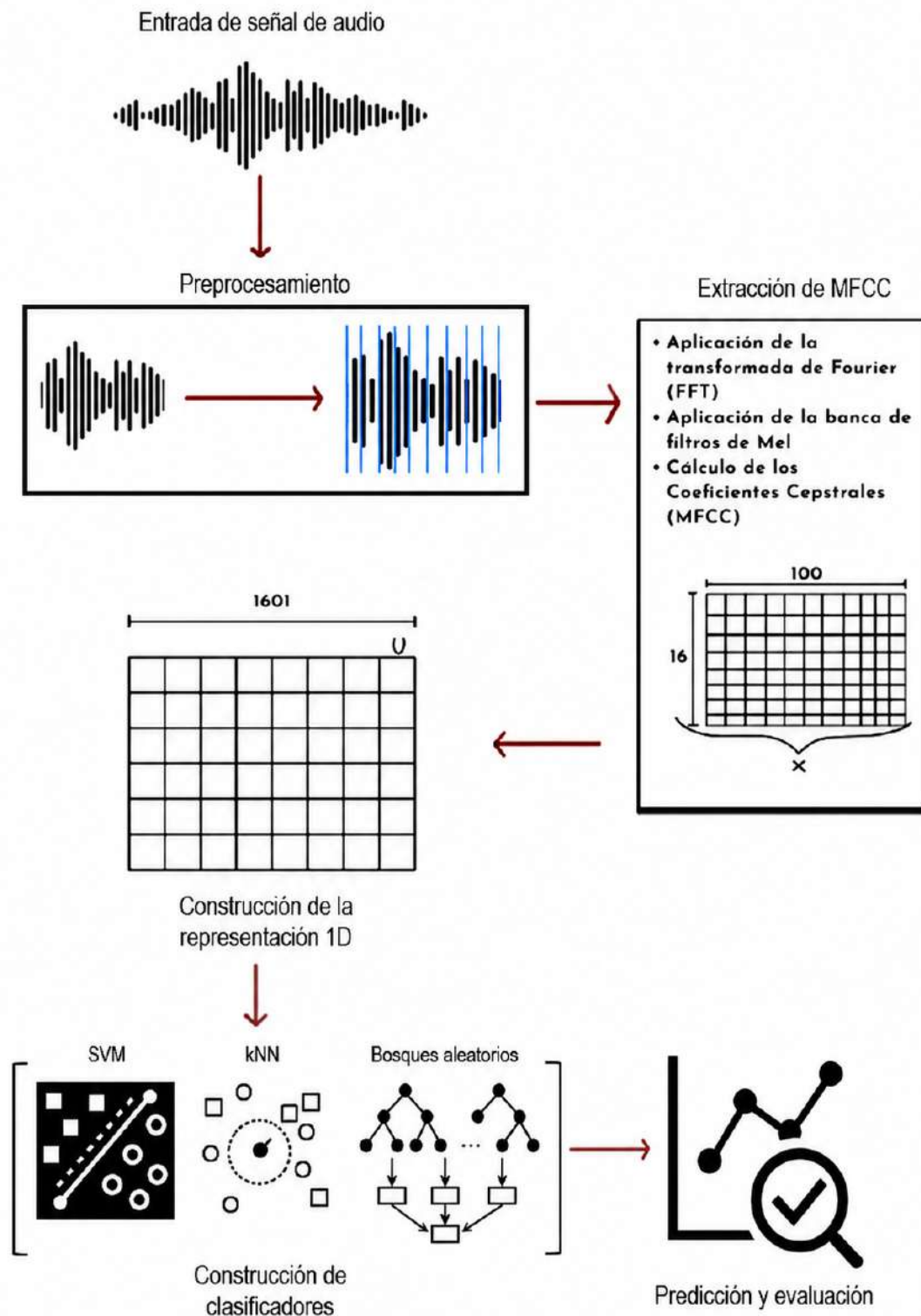


Figura 4.2. Diagrama de flujo de la etapa dos de la metodología propuesta para el procesamiento de señales de audio en 1D, que incluye preprocesamiento, extracción de características MFCC, construcción de representaciones unidimensionales, construcción de clasificadores, clasificación y evaluación.

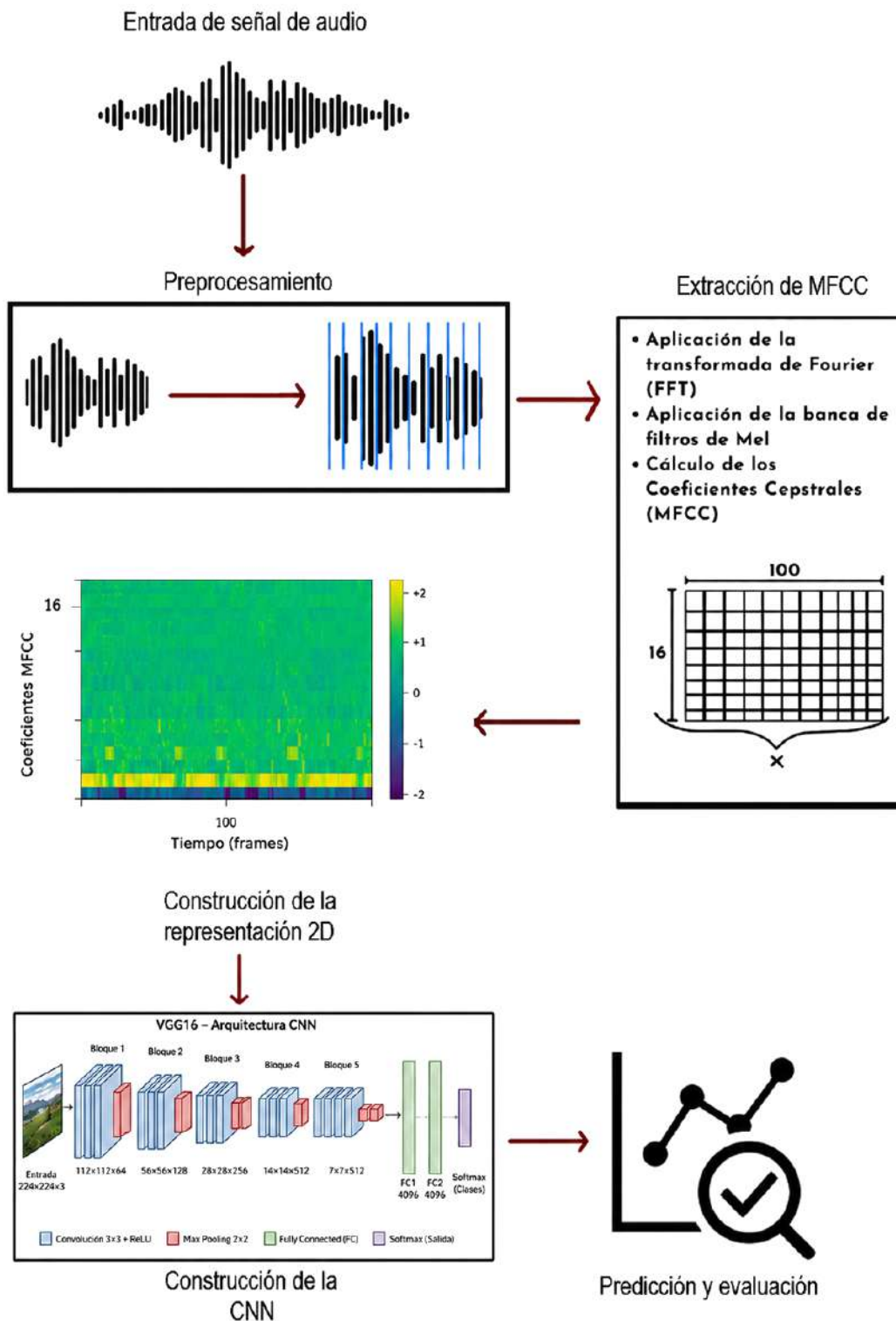


Figura 4.3. Diagrama de flujo de la etapa dos de la metodología propuesta para el procesamiento de señales de audio en 2D, que incluye la extracción de características MFCC, la construcción de representaciones bidimensionales y su clasificación mediante una red neuronal convolucional (VGG16).

4.2.1. Construcción de representaciones

Con la matriz de MFCC generada en la fase anterior, se procedió a elaborar dos tipos de representaciones: una unidimensional (1D) y otra bidimensional (2D). Estas representaciones constituyen la base de los experimentos realizados, pues permiten contrastar el desempeño de clasificadores tradicionales frente a modelos basados en imágenes aplicados a datos acústicos.

Representación unidimensional (1D)

En esta variante, cada matriz de MFCC es transformada en un vector. Para lograrlo, los coeficientes de cada ventana son concatenados en una sola estructura numérica de longitud normalizada, de manera que cada audio es representado como un vector de características de dimensión fija, siendo cada fila un audio. Esta transformación permite adaptar los datos al formato que requieren los clasificadores tradicionales.

Representación bidimensional (2D)

Para este caso, las matrices de MFCC se conservan en su forma original, es decir, como mapas tiempo–frecuencia. Cada matriz se trata como una imagen en una escala de colores, donde el eje horizontal representa el tiempo, el eje vertical corresponde a los coeficientes en la escala Mel y la intensidad de cada punto refleja la energía asociada.

Esta representación permite aprovechar modelos diseñados para el procesamiento de imágenes, como las redes neuronales convolucionales (CNN). Al mantener la estructura bidimensional, las CNN pueden identificar patrones complejos que surgen de la interacción entre variaciones temporales y frecuenciales de la voz, lo que abre la posibilidad de obtener un desempeño superior en tareas de clasificación de hablantes.

4.2.2. Construcción de modelos computacionales de clasificación

Una vez generadas las representaciones 1D y 2D de los MFCC, se procede a la construcción de los modelos de clasificación. El objetivo de esta etapa es entrenar sistemas capaces de identificar a qué hablante corresponde cada muestra de voz, contrastando el desempeño de técnicas tradicionales frente a modelos de aprendizaje profundo.

Clasificación usando representaciones 1D

Para la representación unidimensional, se seleccionan clasificadores tradicionales de aprendizaje automático orientados al manejo de espacios vectoriales planos. En particular, el diseño metodológico incorpora el uso de las máquinas de vectores de soporte (SVM) aplicadas mediante el software Weka, integrándose de forma complementaria algoritmos basados en bosques aleatorios (*Random Forest*) y k -vecinos más cercanos (k -NN) para establecer un entorno de evaluación comparativo amplio. Estos algoritmos son ampliamente utilizados en problemas de clasificación con espacios de características de alta dimensión, como se ha analizado en las secciones anteriores.

El entrenamiento se realiza tomando como entrada los vectores de MFCC correspondientes a cada hablante. Una vez entrenados, los modelos son capaces de realizar la predicción de los hablantes según un vector nuevo de características.

Clasificación usando representaciones 2D

En el caso de la representación bidimensional, se usa una arquitectura VGG16. Las matrices de MFCC se transforman en representaciones visuales que preservan la relación tiempo–frecuencia de la señal. Durante esta etapa, el modelo procesa las imágenes generadas a partir de los MFCC con el objetivo de identificar patrones espaciales asociados a las características de la voz de cada hablante.

A partir de este procesamiento, el modelo aprende a distinguir entre las diferentes clases correspondientes a cada hablante, permitiendo realizar la predicción de la identidad del hablante a

partir de nuevas muestras de audio representadas en este mismo formato bidimensional.

4.2.3. Estrategia del bloque experimental

Con el propósito de evaluar las variables que influyen en el reconocimiento de hablantes, el desarrollo metodológico se estructuró en una secuencia de experimentos específicos. Estas pruebas permiten analizar el efecto individual del escalamiento de variables, la selección de algoritmos y dimensiones vectoriales, la adición de características dinámicas delta y la reducción de atributos por ganancia de información. Asimismo, el bloque experimental contempla el análisis del contenido verbal común y variado para finalmente contrastar dichos resultados frente al rendimiento de la clasificación convolucional 2D bajo diferentes densidades espectrales y cantidades de clases.

La última etapa de la metodología consistió en la evaluación del desempeño de los clasificadores construidos. Para ello, se utilizaron métricas estándar en tareas de clasificación supervisada, lo que permitió cuantificar de manera objetiva el grado de acierto de cada modelo. Esta fase se explicará a detalle en el capítulo a continuación.

Capítulo 5

Experimentos y resultados

En este capítulo se presentan los experimentos realizados para evaluar el desempeño de las representaciones de la señal de voz utilizadas en este trabajo. Las pruebas se organizaron en distintas etapas con el objetivo de analizar cómo influyen diversos factores en la capacidad de reconocimiento de hablantes, tales como el tipo de clasificador, la cantidad de coeficientes cepstrales en la escala de Mel (MFCC, por sus siglas en inglés: Mel Frequency Cepstral Coefficients) empleados y la representación de las características en formato unidimensional o bidimensional.

5.1. Configuración general

Para el presente trabajo se utilizó una configuración estándar sobre los audios y el preprocesamiento:

- Tres segundos por audio: los audios son recortados a tres segundos desde el centro, es decir, 1.5 segundos a la izquierda y a la derecha desde el centro, con el objetivo de mantener una consistencia en todos los audios.
- Relleno: para los audios que no alcanzaran la duración mínima, se aplicó relleno con ceros en los espacios faltantes hasta alcanzar la longitud deseada.
- Fragmentos de los conjuntos de datos: para los experimentos se emplearon los conjuntos de datos de DIMEX y VoxCeleb. Ambos conjuntos de datos se limitaron a 50 personas seleccionadas con 51 audios cada una, dando un total de 2550 instancias, explicado en la sección 3.2.3.

Con respecto a los clasificadores, para el desarrollo de las pruebas basadas en representaciones unidimensionales se seleccionaron tres algoritmos tradicionales de aprendizaje automático

con la finalidad de evaluar de forma comparativa su capacidad de discriminación, realizándose la implementación formal en el software Weka y la calibración de sus estructuras internas a partir de los siguientes parámetros de diseño:

- Máquinas de vectores de soporte (SVM, por sus siglas en inglés: *Support Vector Machines*): se empleó el algoritmo de optimización mínima secuencial (SMO, por sus siglas en inglés: *Sequential Minimal Optimization*) configurado con una función de kernel polinomial de exponente igual a uno, determinándose un parámetro de complejidad $C = 1,0$, un valor de tolerancia de 0.001 y un umbral épsilon de corte de 1.0E-12.
- Bosques aleatorios (*Random Forest*): se aplicó el clasificador basado en árboles de decisión configurando el parámetro *numTrees* en 100 para establecer un ensamble de cien árboles independientes, manteniéndose el parámetro *maxDepth* en cero para permitir un crecimiento ilimitado en la profundidad de los nodos y el parámetro *numFeatures* en cero para la selección automática de atributos discriminantes.
- k -vecinos más cercanos (k -NN, por sus siglas en inglés: *k-Nearest Neighbors*): se implementó el algoritmo IBk parametrizado bajo el valor de KNN igual a ocho para la evaluación de ocho vecinos cercanos, seleccionándose la opción de ausencia de ponderación por distancia y el método de búsqueda lineal fundamentado en la métrica de distancia euclidiana estándar para determinar la proximidad espacial de los vectores.

Con la finalidad de analizar el impacto del escalamiento de los datos en el rendimiento de los modelos vectoriales, las características acústicas se prepararon bajo tres condiciones iniciales que incluyeron la ausencia de procesamiento, la normalización absoluta de los atributos y la estandarización interna de las variables. Por otra parte, para los experimentos usando representaciones 2D se empleó una red neuronal convolucional basada en la arquitectura VGG16 preentrenada, lo que significa que el modelo fue adaptado para el problema de reconocimiento de hablantes ajustando las capas finales para la clasificación multiclase.

Para la evaluación se tomaron en consideración la exactitud (proporción de instancias que se clasificaron correctamente respecto al total de instancias) y el puntaje F1 macro (métrica que permite evaluar el desempeño promedio al considerar todas las clases de forma equilibrada). Además, se analizaron la precisión y la sensibilidad en algunos experimentos para obtener una interpretación un poco más clara del comportamiento del modelo.

En cuanto a los experimentos 1D, se utilizó una validación cruzada de 10 pliegues para evaluar el desempeño. Para los experimentos 2D se realizó una separación de los datos en proporción de 90 % para entrenamiento y 10 % para prueba, en donde los datos fueron separados aleatoriamente según esta proporción. Los experimentos se detallarán a continuación.

5.2. Experimentos usando representaciones 1D

5.2.1. Experimento 1. Comparación del impacto de la normalización y estandarización en las características MFCC

El experimento tiene la finalidad de determinar si el procesamiento de los MFCC aplicando normalización o estandarización tiene un impacto positivo en el rendimiento de la clasificación con SVM comparado con el uso de las características sin procesar. Para ello, las características MFCC fueron evaluadas en tres condiciones: sin procesamiento, normalizadas y estandarizadas.

Como se aprecia en la Tabla 5.1, los resultados muestran que el mejor tipo de preprocesamiento es la estandarización. Los datos sin procesar obtuvieron rendimientos bastante regulares, mientras que la normalización produjo resultados satisfactorios. Por su parte, la estandarización logró los mejores resultados en ambos conjuntos de datos, manteniéndose constante. Estos hallazgos respaldan la elección de aplicar este tipo de preprocesamiento para optimizar el rendimiento del clasificador SVM.

Tabla 5.1. Métricas promedio por conjunto de datos y método de preprocesamiento

Dataset	Procesamiento	Exactitud	Macro F1
DIMEX	Ninguno	45.80 %	45.90 %
	Normalización	86.90 %	86.90 %
	Estandarización	88.08 %	88.18 %
VoxCeleb	Ninguno	22.2 %	22.7 %
	Normalización	37.3 %	37.1 %
	Estandarización	37.60 %	37.33 %

5.2.2. Experimento 2. Evaluación de diferentes algoritmos de clasificación

El objetivo de este experimento es determinar el algoritmo de clasificación más adecuado, así como el número óptimo de coeficientes, para la tarea de identificación de hablantes. Para ello, se realiza una evaluación comparativa entre tres algoritmos ampliamente utilizados en problemas de clasificación supervisada: máquinas de vectores de soporte, bosques aleatorios y k -vecinos más cercanos. Asimismo, se evaluaron distintas configuraciones en el número de coeficientes MFCC (12 y 16 son números comunes en configuraciones de MFCC, mientras que actualmente se exploran cantidades más elevadas como 100 y 200) con el propósito de analizar cómo la dimensionalidad de la representación acústica influye en el desempeño de los clasificadores. Los resultados se comparan utilizando las métricas de exactitud y macro F1.

Los resultados se muestran en la Tabla 5.2. Se observa que el algoritmo de SVM obtuvo el mejor desempeño en todas las métricas. Tanto bosques aleatorios como k -NN obtuvieron resultados bastante deficientes; sin embargo, este último logró alcanzar algunos resultados decentes. La evidencia justifica la elección de emplear SVM como el algoritmo base que se usará en los siguientes experimentos para probar la integridad de los datos.

Tabla 5.2. Resultados de la evaluación de clasificadores bajo la variación en la cantidad de coeficientes MFCC

Dataset	Coef.	Clasificador	Exactitud	Macro F1
DIMEX	12	SVM	76.78 %	77.17 %
		Bosques aleatorios	25.18 %	22.87 %
		k -NN (k=8)	26.55 %	24.84 %
	16	SVM	88.08 %	88.18 %
		Bosques aleatorios	27.88 %	26.89 %
		k -NN (k=8)	40.43 %	38.59 %
	100	SVM	94.16 %	93.78 %
		Bosques aleatorios	20.39 %	17.93 %
		k -NN (k=8)	7.84 %	6.05 %
	200	SVM	92.55 %	92.60 %
		Bosques aleatorios	15.14 %	14.70 %
		k -NN (k=8)	5.13 %	2.50 %
VoxCeleb	12	SVM	37.61 %	37.33 %
		Bosques aleatorios	14.98 %	14.52 %
		k -NN (k=8)	12.35 %	11.04 %
	16	SVM	49.14 %	50.44 %
		Bosques aleatorios	15.92 %	15.15 %
		k -NN (k=8)	16.27 %	15.12 %
	100	SVM	58.71 %	55.92 %
		Bosques aleatorios	7.05 %	6.90 %
		k -NN (k=8)	7.56 %	5.13 %
	200	SVM	54.62 %	54.40 %
		Bosques aleatorios	4.43 %	2.40 %
		k -NN (k=8)	5.13 %	5.00 %

Por otro lado, la comparación entre configuraciones de coeficientes MFCC revela una tendencia clara. Como se ve en la Tabla 5.2, el rendimiento de los clasificadores mejora progresivamente a medida que se incrementa el número de coeficientes, especialmente entre 12 y 100. Sin embargo, al superar este umbral, el desempeño comienza a disminuir ligeramente, lo que sugiere la existencia de un punto óptimo en torno a los 100 coeficientes. Este comportamiento indica que un aumento excesivo en la dimensionalidad puede introducir ruido o redundancia, afectando negativamente los resultados.

5.2.3. Experimento 3. Evaluación del uso de deltas como características de la representación

Este experimento evalúa si la adición de coeficientes delta a las características previas de MFCC provoca un rendimiento mayor para el clasificador SVM en la tarea de reconocimiento de hablantes.

En general, los resultados de la Tabla 5.3 señalan que, en ambos conjuntos de datos, la inclusión de los coeficientes delta no mejora el desempeño del clasificador. De forma más detallada, se nota la disminución de todas las métricas de rendimiento. Esto significaría que la incorporación de estas características, aunque representa conocimiento adicional, no se traduce en información relevante, ya sea por la duración de los segmentos o por el incremento de la dimensionalidad.

Tabla 5.3. Métricas promedio por representación

Dataset	Representación	Exactitud (%)	Macro F1-score
DIMEX	MFCC	94.16 %	94.30 %
	MFCC + deltas	76.31 %	77.70 %
VoxCeleb	MFCC	58.70 %	59.00 %
	MFCC + deltas	38.62 %	39.40 %

Teniendo en cuenta los resultados del experimento, para las siguientes pruebas se decidió continuar utilizando únicamente las características MFCC estandarizadas, que demostraron ofrecer el mejor balance entre precisión y simplicidad del modelo.

5.2.4. Experimento 4. Uso de selección de características

Este experimento tiene como objetivo evaluar si la selección de atributos relevantes mediante la métrica de ganancia de información mejora el rendimiento del clasificador SVM al trabajar con características MFCC estandarizadas. A diferencia de técnicas de reducción de dimensionalidad como el análisis de componentes principales (PCA, por sus siglas en inglés: *Principal Component Analysis*), la ganancia de información permite identificar y conservar únicamente aquellas características que aportan mayor valor discriminativo para la tarea de clasificación.

Para ello, se utilizó un clasificador basado en máquinas de soporte vectorial (SVM), acompañado de un esquema de preprocesamiento que incluye la estandarización de atributos y la selección de características mediante ganancia de información. Esta configuración asegura que, en cada partición de la validación cruzada, tanto la estandarización como la selección se apliquen únicamente sobre los datos de entrenamiento, evitando sesgos y garantizando una evaluación justa del rendimiento.

Aunque la aplicación de selección de características mediante ganancia de información permitió reducir significativamente el número de atributos, se observó una disminución considerable en la exactitud, como se ve en la Tabla 5.4 y en la Figura 5.1. Esto indica que, si bien el método logra conservar las variables con más información, también puede descartar atributos que, aunque menos relevantes individualmente, contribuyen de forma complementaria a la discriminación entre hablantes.

Tabla 5.4. Resultados de reducción de atributos en el conjunto de datos DIMEX

Datos	Atributos seleccionados	Exactitud
DIMEX	2000 atributos	82.43 %
	5000 atributos	87.73 %
	8000 atributos	91.76 %
	Original (10000 atributos)	94.16 %

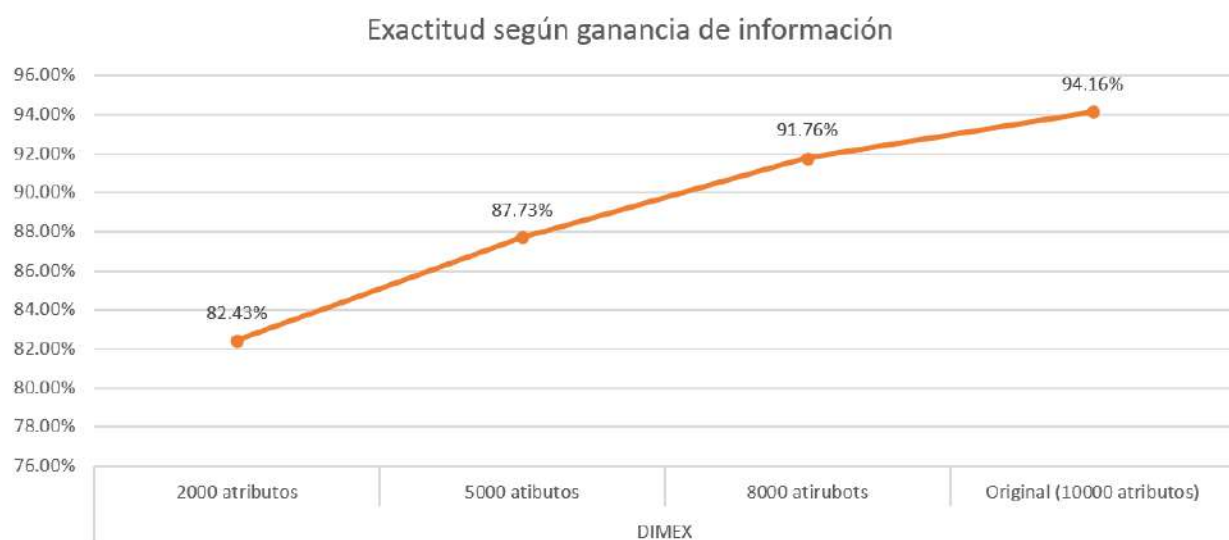


Figura 5.1. Comparación de la exactitud entre datos originales y reducidos

Estos resultados sugieren que, en contextos donde se requiere preservar la máxima capacidad discriminativa, como en las tareas de reconocimiento de hablantes, una reducción agresiva del espacio de atributos puede ser contraproducente. Por tanto, si el fin es trabajar con conjuntos de datos más extensos y complejos, será necesario evaluar cuidadosamente el equilibrio entre eficiencia computacional y pérdida de información antes de aplicar técnicas de selección.

5.2.5. Experimento 5. Audios comunes vs individuales (DIMEX)

El conjunto de datos DIMEX presenta una característica distintiva: sus grabaciones se dividen en dos categorías. Por un lado, se encuentran las frases comunes, en las que todos los participantes pronuncian exactamente las mismas oraciones, y por otro, las frases individuales, donde cada hablante emite un conjunto único de 50 frases, distinto e irrepetible entre participantes.

Este experimento tiene como propósito evaluar el impacto del tipo de contenido (controlado o variado) sobre el rendimiento del modelo de clasificación. En particular, se busca determinar si el uso de frases comunes (compartidas entre hablantes) o individuales (propias de cada persona) influye en la capacidad del clasificador para identificar correctamente al hablante.

Los resultados muestran una diferencia notable entre ambos escenarios, como se ve en la Tabla 5.5: cuando los hablantes utilizan frases comunes, el modelo reduce significativamente su rendimiento, mientras que con frases individuales alcanza un desempeño mucho mayor.

Tabla 5.5. Rendimiento del modelo SVM con audios comunes e individuales

Comunes vs individuales	Exactitud (%)	Macro F1
Comunes	72.11 %	70.90 %
Individuales	94.16 %	94.20 %

Esto sugiere que el contenido verbal influye directamente en la variabilidad de las características acústicas capturadas. Las frases diferentes permiten al modelo abarcar un rango más amplio de información prosódica y espectral, facilitando una mejor discriminación entre hablantes.

5.3. Experimentos usando representaciones 2D

5.3.1. Experimento 6. Representación bidimensional (imágenes de MFCC) con tres clases

En este experimento se empleó una representación bidimensional de los coeficientes MFCC para transformar cada muestra de voz en una imagen que preserva la variación temporal y frecuencial de la señal. Esta estructura permite aprovechar el potencial de las redes neuronales convolucionales (CNN, por sus siglas en inglés: *Convolutional Neural Networks*) en el reconocimiento de patrones espaciales, explorando si el modelo identifica diferencias entre hablantes a partir de mapas espectrales. Para ello, se seleccionó un subconjunto aleatorio y balanceado de tres hablantes (u02, u03 y u13) para las fases de entrenamiento, validación y prueba, construyéndose el sistema sobre una arquitectura *VGG16* preentrenada mediante transferencia de aprendizaje en sus capas superiores.

El modelo alcanzó una exactitud promedio del 63 %, apreciándose un comportamiento heterogéneo al analizar el desempeño numérico por categoría. Como se evidencia en la Tabla 5.6, la clase u02 obtuvo una precisión perfecta del 100 %, lo que demuestra la ausencia de falsos positivos; sin embargo, registró la sensibilidad o recall más deficiente con un 10 %, indicándose que el clasificador subestimó esta categoría al asignar erróneamente el 90 % de sus muestras hacia otras clases.

Tabla 5.6. Métricas promedio para la representación bidimensional con tres clases y 100 coeficientes MFCC

Clase	Precisión	Recall	F1-score
u02	1.00	0.10	0.18
u03	1.00	0.80	0.89
u13	0.48	1.00	0.65
Promedio macro	0.83	0.63	0.57

Por el contrario, la clase u13 expuso un escenario inverso al alcanzarse un recall máximo del 100 %, lo que asegura la identificación de todas sus instancias reales a costa de una reducción

en la precisión hasta el 48 %. Esta asimetría revela un sesgo de predicción hacia el hablante u13, catalogándose muestras ajenas dentro de esta categoría e incrementando los falsos positivos. Finalmente, la clase u03 reflejó el rendimiento más equilibrado del entorno con puntuaciones del 100 % en precisión y del 80 % en recall, consolidándose el puntaje F1-score más alto con un 89 %.

Este comportamiento se explica por la complejidad estructural de las redes profundas y la alta sensibilidad de la arquitectura VGG16 ante conjuntos de datos reducidos. Al disponer de un volumen limitado de datos, las capas convolucionales presentan dificultades para generalizar patrones robustos de textura y frecuencia, asimilándose características que desbalancean los criterios de decisión en la etapa final. No obstante, la respuesta general del sistema aporta evidencia valiosa sobre la viabilidad del enfoque visual, confirmándose que las imágenes de los coeficientes MFCC preservan rasgos acústicos discriminativos con potencial de mejora mediante la ampliación del corpus.

5.3.2. Experimento 7. Representación bidimensional con cinco clases

Este experimento repite la base del anterior, pero con una complejidad aumentada. El objetivo es analizar cómo se comporta la CNN en una comparación multiclase con más hablantes. Para ello se empleó igualmente la representación bidimensional de los coeficientes MFCC, ahora aplicada a cinco personas. Se utilizó la misma arquitectura de red VGG16.

Tabla 5.7. Métricas promedio para la representación bidimensional con cinco clases y 100 coeficientes MFCC

Clase	Precisión	Recall	F1-score
s024	1.00	1.00	1.00
s033	1.00	1.00	1.00
s038	0.71	0.83	0.77
s042	0.80	0.67	0.73
s044	1.00	1.00	1.00
Promedio macro	0.90	0.90	0.90

El modelo alcanzó una exactitud de 90 % en el conjunto de prueba, mostrando un desempeño sólido incluso al incrementar el número de clases respecto al experimento anterior. Tres de

los cinco hablantes fueron clasificados sin error, mientras que la principal dificultad se concentró en las clases s038 y s042, donde se observaron confusiones mutuas.

Este comportamiento sugiere que la representación bidimensional de 100 coeficientes MFCC conserva suficiente información espectral para que la red neuronal aprenda patrones útiles de discriminación entre hablantes. No obstante, el hecho de que las clases s038 y s042 presenten menor precisión y recall indica que existen similitudes acústicas entre estos hablantes, lo que dificulta su separación completa. Aun así, el resultado obtenido confirma que el enfoque basado en imágenes MFCC y redes convolucionales puede alcanzar un desempeño competitivo en escenarios multiclase de mayor complejidad.

5.3.3. Experimento 8. Representación bidimensional con cinco clases y 40 coeficientes MFCC

Se realizó un experimento adicional utilizando 40 coeficientes MFCC en lugar de 100 con el fin de ver si una amplia cantidad de coeficientes introducía ruido en las muestras. La elección de este valor respondió a la necesidad de reducir la cantidad de información de entrada sin llevarla a un punto excesivamente bajo que provocara una pérdida abrupta de características relevantes de la voz. Por lo tanto, 40 coeficientes se consideraron un límite intermedio suficiente para conservar una representación descriptiva de la señal y, al mismo tiempo, permitir evaluar si una configuración más extensa, como la de 100 coeficientes, introducía información redundante o ruido.

Para este experimento se utilizó el mismo subconjunto de cinco hablantes empleado en la configuración anterior, manteniendo la división en entrenamiento, validación y prueba.

El modelo alcanzó una exactitud de 86.67 %, con un promedio macro de 0.86 en F1-score. En comparación con el experimento realizado con 100 coeficientes MFCC, se observa una disminución en el desempeño, ya que la configuración anterior con 100 coeficientes obtuvo una exactitud de 90 %. Esta diferencia parece indicar que, en el caso de la representación bidimensional, la reducción a 40 coeficientes implica una pérdida de información útil para la discriminación entre hablantes.

Tabla 5.8. Métricas promedio para la representación bidimensional con cinco clases y 40 coeficientes MFCC

Clase	Precisión	Recall	F1-score
s024	0.86	1.00	0.92
s033	1.00	0.83	0.91
s038	1.00	0.50	0.67
s042	0.75	1.00	0.86
s044	0.86	1.00	0.92
Promedio macro	0.89	0.87	0.86

Los resultados parecen apuntar a que la configuración con 100 coeficientes MFCC resulta más adecuada para este escenario, ya que conserva una mayor cantidad de información discriminativa sin llegar a afectar negativamente el proceso de clasificación.

5.4. Comparación de 1D y 2D

Esta sección presenta una comparación entre las representaciones unidimensionales y bidimensionales de los coeficientes MFCC para la tarea de identificación de hablantes. Para ello se replicaron en formato unidimensional los mismos conjuntos de datos utilizados en los experimentos bidimensionales, permitiendo contrastar el desempeño de ambos enfoques bajo condiciones experimentales equivalentes. En concreto, se realizaron los siguientes experimentos:

- Experimento con tres hablantes.
- Experimento con cinco hablantes utilizando 100 coeficientes MFCC.
- Experimento con cinco hablantes utilizando 40 coeficientes MFCC.

En general, los resultados obtenidos permiten comparar directamente el comportamiento de ambas representaciones manteniendo constantes las condiciones del conjunto de datos. Los experimentos se detallan en las siguientes secciones:

Comparación para tres hablantes

Con el objetivo de analizar el efecto del tipo de representación en el desempeño del sistema de identificación de hablantes, se realizó una comparación entre las representaciones unidimensionales y bidimensionales de los coeficientes MFCC utilizando el mismo subconjunto de tres hablantes del experimento seis en el apartado 5.3.1.

En ambos casos se empleó el mismo conjunto de datos. La representación unidimensional utilizó vectores MFCC como entrada para un clasificador de máquinas de vectores de soporte (SVM), mientras que la representación bidimensional transformó los coeficientes MFCC en imágenes procesadas mediante una red convolucional basada en *VGG16*.

La Tabla 5.9 muestra las métricas promedio obtenidas para ambas representaciones.

Tabla 5.9. Comparación entre representaciones 1D y 2D para el experimento con tres hablantes

Representación	Exactitud global	Precisión promedio	Recall promedio	F1-score promedio
MFCC 1D (SVM)	0.83	0.84	0.83	0.83
MFCC 2D (CNN)	0.63	0.83	0.63	0.57

Los resultados indican que la representación unidimensional obtiene un mejor desempeño en este escenario, alcanzando una exactitud global del 83 % y un F1-score promedio de 0.83. Por su parte, la representación bidimensional obtiene una exactitud global del 63 % y un F1-score promedio de 0.57.

Esta diferencia puede explicarse principalmente por el tamaño reducido del conjunto de datos. Las redes convolucionales suelen requerir una mayor cantidad de muestras para aprender representaciones robustas, mientras que modelos basados en vectores como SVM pueden adaptarse mejor a conjuntos pequeños de datos utilizando descriptores acústicos compactos.

Este resultado sirve como punto de referencia inicial para las comparaciones posteriores con conjuntos de mayor tamaño y diferentes configuraciones de coeficientes MFCC.

Comparación para cinco hablantes (100 MFCC)

Para ampliar el análisis realizado en la comparación anterior, se llevó a cabo una comparación adicional entre las representaciones unidimensionales y bidimensionales utilizando el conjunto de cinco hablantes del experimento siete en el apartado 5.3.2.

La Tabla 5.10 muestra las métricas promedio obtenidas para ambos métodos.

Tabla 5.10. Comparación entre representaciones 1D y 2D para el experimento con cinco hablantes y 100 MFCC

Representación	Exactitud global	Precisión promedio	Recall promedio	F1-score promedio
MFCC 1D (SVM)	0.96	0.96	0.96	0.96
MFCC 2D (CNN)	0.90	0.90	0.90	0.90

Los resultados indican que la representación unidimensional alcanza un desempeño superior en este experimento, obteniendo una exactitud global del 96 % y un F1-score promedio de 0.96. Por su parte, la representación bidimensional alcanza una exactitud global del 90 % y un F1-score promedio de 0.90.

En consecuencia, para este conjunto de datos y esta configuración de características, la representación unidimensional basada en MFCC indica ser ligeramente más efectiva que la representación bidimensional. No obstante, al comparar estos resultados con el experimento de tres hablantes se observa una mejora considerable en el desempeño y equilibrio de las clases del modelo bidimensional, lo que sugiere que las redes convolucionales pueden beneficiarse de conjuntos de datos más amplios para aprender patrones más robustos.

Comparación para cinco hablantes (40 MFCC)

Para completar el análisis comparativo entre representaciones unidimensionales y bidimensionales, se evaluó el mismo conjunto de cinco hablantes utilizando una representación basada en 40 coeficientes MFCC, con la misma configuración del experimento ocho en el apartado 5.3.3.

La Tabla 5.11 muestra las métricas promedio obtenidas para ambos métodos.

Tabla 5.11. Comparación entre representaciones 1D y 2D para el experimento con cinco hablantes y 40 MFCC

Representación	Exactitud global	Precisión promedio	Recall promedio	F1-score promedio
MFCC 1D (SVM)	0.97	0.97	0.97	0.97
MFCC 2D (CNN)	0.87	0.89	0.87	0.86

Los resultados indican nuevamente un mejor desempeño para la representación unidimensional, la cual alcanza una exactitud global del 97 % y un F1-score promedio de 0.97. Por su parte, la representación bidimensional obtiene una exactitud global del 87 % y un F1-score promedio de 0.86.

En comparación con el experimento realizado utilizando 100 coeficientes MFCC, se observa que la reducción a 40 coeficientes no afecta negativamente el desempeño del modelo basado en SVM, mientras que en el caso de la representación bidimensional sí se observa una ligera disminución en el rendimiento. Esto sugiere que el clasificador SVM mantiene un desempeño estable incluso con representaciones más compactas de los MFCC.

5.5. Análisis de resultados y hallazgos clave

A lo largo del desarrollo se aplicó una secuencia de pruebas orientadas a identificar los factores que influyen en la precisión del modelo y en la interpretabilidad de los resultados.

Se comprobó que las etapas de normalización y estandarización de los coeficientes influyen de manera significativa en la estabilidad de los clasificadores. La estandarización permitió reducir el sesgo introducido por la variabilidad entre grabaciones, estabilizando los datos.

La comparación entre distintos clasificadores tradicionales (SVM, bosques aleatorios y k -NN) y diferentes cantidades de coeficientes MFCC (12, 16, 100 y 200) evidenció que el modelo SVM ofrece el mejor equilibrio entre rendimiento y generalización. En el conjunto DIMEX, el desempeño óptimo se alcanzó con 100 coeficientes, obteniendo una exactitud promedio de 94.16 %, mientras que en VoxCeleb el resultado fue de 58.7 %, lo que refleja la mayor variabilidad de condiciones acústicas presentes en dicho conjunto de datos. Esto apunta a que, en contextos con menor ruido y homogeneidad de grabaciones, el enfoque unidimensional es altamente efectivo.

El tercer experimento permitió evaluar la inclusión de coeficientes delta, con el objetivo de capturar variaciones temporales en la señal. Los resultados demostraron que su incorporación no incrementa el rendimiento del modelo, e incluso puede introducir ruido o redundancia. Esto sugiere que la información presente en los MFCC originales ya es suficiente para representar la estructura temporal relevante de los hablantes, al menos bajo las condiciones utilizadas.

Posteriormente, se aplicó una técnica de selección de atributos para analizar el impacto de la reducción dimensional. Si bien se observó que la disminución progresiva del número de atributos no provoca una caída drástica en el rendimiento, sí afecta de manera perceptible la precisión del modelo. De manera particular, al conservar únicamente una fracción de los atributos originales, se pierde información relevante para la discriminación entre hablantes, reduciendo la exactitud hasta en un 12 % en la prueba más alta. Por esta razón, se determinó que la reducción de características mediante ganancia de información no resulta adecuada dentro del marco experimental propuesto, donde el objetivo principal es preservar la mayor cantidad posible de información acústica para

asegurar comparaciones fieles entre representaciones.

Otro aspecto explorado fue el impacto del contenido verbal en el reconocimiento. Se demostró que el uso de frases distintas por hablante genera modelos más robustos, alcanzando una exactitud de 94.16 % frente al 72.11 % obtenido con frases comunes. Esto sugiere que la variabilidad del contenido en las oraciones permite capturar un rango más amplio de características de la voz del individuo, mientras que las frases controladas tienden a limitar la diversidad de patrones acústicos que el modelo puede aprender, lo que, al final, dificulta la capacidad del modelo para discriminar hablantes.

Finalmente, se llevaron a cabo tres experimentos comparativos destinados a comparar la representación unidimensional con la representación bidimensional utilizando el mismo marco experimental y los mismos subconjuntos de hablantes. En estos experimentos se evaluaron tres configuraciones: un escenario reducido con tres hablantes, y dos configuraciones con cinco hablantes empleando representaciones basadas en 100 y 40 coeficientes MFCC, respectivamente.

En el escenario de tres hablantes, el clasificador SVM aplicado sobre vectores MFCC alcanzó una exactitud del 83.33 %, mientras que la red neuronal convolucional aplicada sobre las imágenes de MFCC obtuvo un 63 %. Esta diferencia evidencia la mayor estabilidad del enfoque vectorial en conjuntos de datos pequeños y controlados, así como la alta sensibilidad que tienen los modelos convolucionales a la cantidad de datos disponibles.

Cuando el número de hablantes se incrementó a cinco y se continuaron utilizando 100 coeficientes MFCC, ambos enfoques mostraron un desempeño elevado, aunque el modelo SVM continuó presentando una ligera ventaja, alcanzando una exactitud del 96 % frente al 90 % obtenido por la red convolucional. Este resultado sugiere que, bajo condiciones de mayor diversidad de datos, los modelos basados en aprendizaje profundo comienzan a aproximarse al desempeño de los clasificadores tradicionales.

Al repetir el experimento con cinco hablantes utilizando únicamente 40 coeficientes MFCC, se observó que el clasificador SVM mantiene un rendimiento estable, alcanzando una exactitud del 97 %, mientras que el modelo bidimensional presenta una ligera disminución hasta un 86 %. Este

comportamiento indica que la reducción en el número de coeficientes no afecta de forma significativa al modelo vectorial, pero sí puede reducir la cantidad de información espectral disponible para las redes convolucionales.

A partir del conjunto de resultados obtenidos, se puede concluir que la representación unidimensional basada en MFCC constituye, bajo las condiciones experimentales analizadas, una opción estable, eficiente y confiable para el reconocimiento de hablantes. Esta afirmación parte de la secuencia de experimentos que incluyó la evaluación de técnicas de normalización, la variación en el número de coeficientes, la incorporación de deltas, la selección de atributos por ganancia de información y el análisis del contenido verbal. Cada uno de estos aportó información sobre cómo se comporta el modelo ante distintas configuraciones, consolidando el enfoque 1D como una solución robusta en contextos con conjuntos de datos pequeños o moderados.

Por otro lado, la representación bidimensional mediante mapas de calor de MFCC y redes convolucionales, aunque su rendimiento fue inferior al enfoque 1D en todos los escenarios evaluados (63 % frente a 83 % en el experimento con tres hablantes, y 90 % frente a 96 % en la configuración con cinco hablantes y 100 MFCC), reveló aspectos clave sobre la sensibilidad de estos modelos a la cantidad de datos, la arquitectura y la calibración de hiperparámetros.

Asimismo, al analizar los experimentos con cinco hablantes se observa que el desempeño del modelo bidimensional mejora considerablemente respecto al escenario reducido, pasando de una exactitud del 63 % a valores cercanos al 90 %. Este comportamiento sugiere que las redes convolucionales se benefician de una mayor cantidad de datos para aprender representaciones más robustas. Sin embargo, al reducir el número de coeficientes MFCC de 100 a 40 se observa una ligera disminución en el rendimiento del modelo 2D (de 90 % a aproximadamente 86 %), lo que indica que estas arquitecturas dependen en mayor medida de representaciones espectrales más ricas para mantener su capacidad de discriminación entre hablantes.

Capítulo 6

Conclusiones y trabajo futuro

El presente capítulo reúne las principales aportaciones y resultados alcanzados a lo largo del desarrollo de este trabajo de tesis, cuyo propósito fue evaluar y comparar distintas representaciones de la señal de voz, tanto unidimensionales como bidimensionales, aplicadas a la tarea de reconocimiento de hablantes. El estudio se estructuró bajo una metodología experimental inspirada en el proceso KDD (*Knowledge Discovery in Databases*) adaptado al contexto del procesamiento de voz. Dicha metodología permitió organizar las fases del presente trabajo que abarcaron la selección de datos, el preprocesamiento, la extracción de características, la construcción de representaciones y la consolidación de los modelos de clasificación.

6.1. Conclusiones

El presente trabajo de tesis tuvo como propósito evaluar el desempeño de distintas representaciones de la señal de voz en la tarea de reconocimiento de hablantes, considerándose tanto enfoques tradicionales unidimensionales basados en coeficientes cepstrales en la escala de Mel (MFCC, por sus siglas en inglés: *Mel Frequency Cepstral Coefficients*), como una aproximación bidimensional derivada de los mismos coeficientes. Después de la serie de experimentos realizados, se lograron identificar factores críticos que influyen directamente en el rendimiento y en la estabilidad general de los modelos.

A continuación, se exponen los hallazgos principales de la investigación:

- **Impacto del escalamiento estadístico:** La estandarización de las características MFCC demostró ser una etapa de preprocesamiento indispensable, lográndose reducir las diferencias de volumen y energía entre los hablantes causadas por las condiciones de grabación de los audios, lo cual incrementó la estabilidad y la precisión de los clasificadores.

- **Robustez del clasificador base:** El clasificador de máquinas de vectores de soporte (SVM, por sus siglas en inglés: *Support Vector Machines*) se consolidó como el algoritmo tradicional más fuerte y consistente para la tarea de reconocimiento frente a bosques aleatorios y k -vecinos más cercanos, alcanzándose los niveles de exactitud más elevados y demostrando un comportamiento estable al incrementar el número de personas evaluadas.
- **Redundancia de los coeficientes dinámicos:** La inclusión de coeficientes delta provocó una disminución constante en la precisión de los modelos vectoriales, determinándose que para segmentos de voz con una duración fija de tres segundos los MFCC estáticos capturan la información acústica suficiente, mientras que la adición de atributos dinámicos introduce información repetida que causa un sobreajuste en el entrenamiento.
- **Compromiso entre dimensiones y velocidad:** La selección de atributos mediante el criterio de ganancia de información demostró ser una ruta viable para reducir la cantidad de datos, concluyéndose que una pérdida de exactitud cercana al 10 % constituye un intercambio aceptable si el objetivo del sistema prioriza la velocidad de procesamiento en dispositivos con recursos limitados o aplicaciones en tiempo real.
- **Influencia del tipo de frases:** El contenido verbal de las muestras ejerce un efecto directo en la capacidad de acierto del sistema, comprobándose que los modelos entrenados con frases individuales y variadas por hablante poseen mayor precisión que aquellos basados en frases comunes, debido a que la diversidad del texto obliga a los algoritmos a aprender las características físicas de la voz en lugar de memorizar palabras específicas.
- **Cantidad de datos en el contraste 1D frente a 2D:** La comparación entre los enfoques 1D y 2D evidenció que el método unidimensional basado en vectores planos mantiene una estabilidad superior y un alto rendimiento ante conjuntos de datos pequeños o moderados, mientras que el enfoque visual bidimensional depende estrictamente de contar con una gran cantidad de imágenes para que la red neuronal pueda aprender de forma correcta.

- **Sensibilidad a la cantidad de coeficientes:** La reducción en el número de coeficientes MFCC reveló comportamientos opuestos según el tipo de representación, observándose que el modelo SVM unidimensional mantiene su buen desempeño al bajar de 100 a 40 coeficientes, en contraste con la arquitectura VGG16, que sufrió una caída en sus resultados bajo el mismo escenario, lo cual indica que las redes neuronales basadas en imágenes requieren mapas espectrales más completos y ricos en información para poder clasificar correctamente.

La comparación sistemática entre ambos enfoques no solo permitió identificar ventajas y limitaciones técnicas, sino que también evidenció que la elección de la representación debe responder a criterios metodológicos claros como el tipo de modelo, la disponibilidad de datos, la complejidad del problema y los objetivos del sistema. Los resultados respaldan el uso de representaciones unidimensionales como una solución clara, eficiente y fácil de replicar para la tarea de reconocimiento del hablante en escenarios controlados; sin embargo, se reconoce el valor de la aproximación bidimensional, cuya capacidad para analizar la señal de forma visual establece una base sólida para el desarrollo de futuros sistemas de clasificación híbridos.

6.2. Trabajo futuro

Los resultados obtenidos en este trabajo permiten estudiar distintos objetivos que podrían ampliar y profundizar el estudio de las representaciones de la señal de voz en tareas de reconocimiento de hablantes.

Como trabajo futuro, se propone extender los experimentos con representaciones bidimensionales utilizando el conjunto completo de hablantes y fragmentos de audio disponibles en los corpus DIMEX y VoxCeleb. Esta ampliación permitirá evaluar con mayor precisión el potencial de las redes neuronales convolucionales aplicadas a los mapas de coeficientes MFCC. La repetición de las pruebas bajo condiciones más favorables para el aprendizaje profundo, con mayor cantidad de datos, más diversidad de personas y una infraestructura computacional robusta, permitirá determinar si las representaciones 2D pueden igualar o superar el desempeño alcanzado con

los vectores unidimensionales.

Además, se plantea profundizar en el análisis respecto al número de coeficientes empleados, ya que, aunque tradicionalmente se han utilizado 40 MFCC, en la presente investigación se obtuvieron resultados sobresalientes con 100 coeficientes. Si bien otros autores han reportado el uso de esta cantidad, se considera necesario estudiar la resolución espectral óptima para identificar el punto exacto donde el incremento de atributos deja de aportar información útil y comienza a generar redundancia.

Por otro lado, se contempla la posibilidad de investigar enfoques híbridos o multimodales que combinen características derivadas de las representaciones 1D y 2D, integrando de este modo la información temporal y espacial en un mismo modelo de clasificación. Esta fusión podría aportar una descripción más completa de la señal de voz, mejorando la discriminación entre los hablantes e incluso abriendo la posibilidad de ampliar el sistema hacia tareas relacionadas como la verificación de identidad o el reconocimiento de emociones a partir del comportamiento acústico.

Asimismo, se considera relevante investigar técnicas de regularización, aumento de datos y transferencia de aprendizaje que permitan mitigar el sobreajuste en modelos convolucionales cuando se trabaja con conjuntos de datos limitados, como los empleados en este estudio, donde la cantidad de hablantes y audios condicionó el aprendizaje de la red.

Estas proyecciones futuras no solo buscan mejorar el rendimiento general de los modelos, sino también avanzar hacia el desarrollo de sistemas de reconocimiento de voz que resulten más explicables, escalables y adaptables a entornos reales de operación.

Bibliografía

- Ahlawat, S., Choudhary, A., Nayyar, A., Singh, S., & Yoon, B. (2020). Improved Handwritten Digit Recognition Using Convolutional Neural Networks (CNN). *Sensors*, 20(12), 3344. <https://doi.org/10.3390/s20123344>
- Akerberg, M., Espinosa, A., & Santiago, F. (2015). *La enseñanza de la pronunciación: Manual para profesores*. Universidad Nacional Autónoma de México.
- Ali, S., Tanweer, S., Khalid, S. S., & Rao, N. (2020). Mel Frequency Cepstral Coefficient: A Review. *Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development (ICIDSSD 2020)*. <https://doi.org/10.4108/eai.27-2-2020.2303173>
- Alnuaimi, A. F., & Albaldawi, T. H. (2024). An Overview of Machine Learning Classification Techniques [Presented at the Fifth International Scientific Conference of Alkafeel University (ISCKU 2024)]. *BIO Web of Conferences*, 97, 00133. <https://doi.org/10.1051/bioconf/20249700133>
- Alvarado Erazo, E. R. (2016, enero). *Tratamiento digital de la señal de audio utilizando el DSP SHARCK 21061* [Tesis de Ingeniería en Electrónica]. Universidad de San Carlos de Guatemala, Facultad de Ingeniería.
- Alvarado Moya, J. P. (2011). *Procesamiento digital de señales*. Instituto Tecnológico de Costa Rica.
- Belinkov, Y., & Glass, J. R. (2018). Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7, 49-72. <https://api.semanticscholar.org/CorpusID:56657817>
- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.
- Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition [Consultado el 24 de agosto de 2025]. *Data Mining and Knowledge Discovery*, 2(2), 121-167. <https://doi.org/10.1023/A:1009715923555>

- Camargo Abril, G. A. (2024). *Una comparación para el reconocimiento de patrones del habla usando Modelos de Markov Oculto y Redes Neuronales en el idioma Español* [Tesis de maestría]. Universidad Nacional de Colombia. <https://repositorio.unal.edu.co/items/30c96034-ab57-4cbd-8635-6659dda95130>
- Celi-Parraga, R. J., Varela-Tapia, E. A., Acosta-Guzmán, I. L., & Montaña-Pulzara, N. R. (2021). Técnicas de procesamiento de lenguaje natural en la inteligencia artificial conversacional textual. *AlfaPublicaciones*, 3(4.1), 40-52. <https://doi.org/10.33262/ap.v3i4.1.123>
- Chamba, D. A. J., & Jiménez, L. C. (2022). Estudio comprensivo de la Transformada de Fourier Discreta para el análisis de señales digitales. *Revista Científica y Tecnológica UPSE*, 9(1), 75-84. <https://doi.org/10.26423/rctu.v9i1.664>
- Charisma, A., Hidayat, M. R., & Zainal, Y. B. (2017). Speaker recognition using mel-frequency cepstrum coefficients and sum square error. *2017 3rd International Conference on Wireless and Telematics (ICWT)*, 160-163. <https://doi.org/10.1109/ICWT.2017.8284159>
- Chowdhury, A., & Ross, A. (2020). Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals. *IEEE Transactions on Information Forensics and Security*, 15, 1616-1629. <https://doi.org/10.1109/TIFS.2019.2941773>
- Costantini, G., Cesarini, V., & Brenna, E. (2023). High-Level CNN and Machine Learning Methods for Speaker Recognition. *Sensors*, 23(7). <https://doi.org/10.3390/s23073461>
- Departamento de Dispositivos Electrónicos. (2003). Tema 1: Conceptos Básicos de Circuitos Digitales: Señales, sistemas, puertas lógicas y características de familias lógicas [Material Auxiliar de Clase, Dep-Leg. N°: MA-686-2003].
- Díaz, J. (2024). *Comparativa de metodologías de minería de datos: KDD, CRISP-DM y SEMMA explicadas* [Consultado en agosto de 2025]. <https://es.linkedin.com/pulse/comparativa-de-metodolog%C3%ADas-miner%C3%ADa-datos-kdd-crisp-dm-y-semma-explicadas-3uhwe>
- Galindo, E. A., Perdomo, J. A., & Figueroa-García, J. C. (2020). Estudio comparativo entre máquinas de soporte vectorial multiclase, redes neuronales artificiales y sistema de inferencia

- neuro-difuso autoorganizado para problemas de clasificación. *Información Tecnológica*, 31(1), 273-286. <https://doi.org/10.4067/S0718-07642020000100273>
- Gareta Manzano, A. (2015). Extracción de características y clasificación automática de señales de audio.
- González, Y., Juárez, H., Rocha, O., Hernández, R., & Bermúdez, A. (2019). Evaluación comparativa de sistemas de reconocimiento de locutor basados en los algoritmos LPC, CC y MFCC. *Memoria Investigaciones en Ingeniería*, (17), 121-136. <https://doi.org/10.36561/ING.17.6>
- González González, I. (2019, septiembre). *Análisis y comparación de extracción de características en señales de audio* [Trabajo de Fin de Máster]. Escuela Politécnica Superior [Máster Universitario en Ingeniería de Telecomunicación].
- Hegde, S., Achary, K. K., & Shetty, S. (2015). Feature selection using Fisher's ratio technique for automatic speech recognition. <https://arxiv.org/abs/1505.03239>
- Huang, Z. (2024). A Research on Image Recognition and Classification Based on Traditional Machine Learning and Deep Learning. *Transactions on Computer Science and Intelligent Systems Research*, 5, 766-773. <https://doi.org/10.62051/0dbqaa10>
- IBM. (s.f.). *¿Qué es el algoritmo de k-vecinos más cercanos (k-NN)?* [[En línea; accedido el 22-junio-2026]]. <https://www.ibm.com/mx-es/think/topics/knn>
- Jin, S., Wang, X., Du, L., & He, D. (2021). Evaluation and modeling of automotive transmission whine noise quality based on MFCC and CNN. *Applied Acoustics*, 172, 107562. <https://doi.org/10.1016/j.apacoust.2020.107562>
- Liu, X., Sahidullah, M., & Kinnunen, T. (2021). Learnable MFCCs for Speaker Verification. *arXiv preprint arXiv:2102.10322*, 1-5. <https://doi.org/10.48550/arXiv.2102.10322>
- López, R. F., & Fernández, J. M. F. (2008). *Las redes neuronales artificiales*. Netbiblo.
- Mahmood, A., & Köse, U. (2021). Speech recognition based on Convolutional neural networks and MFCC algorithm. 1, 6-12.

- Maia, M., Pimentel, J., Pereira, I., Gondim, J., Barreto, M., & Ara, A. (2020). Convolutional Support Vector Models: Prediction of Coronavirus Disease Using Chest X-rays. *Information*, 11. <https://doi.org/10.3390/info11120548>
- Maimon, O., & Rokach, L. (2010, enero). *Data Mining and Knowledge Discovery Handbook*, 2nd ed.
- Martínez Mascorro, G. A., & Aguilar Torres, G. (2013). Reconocimiento de voz basado en MFCC, SBC y Espectrogramas. *Ingenius. Revista de Ciencia y Tecnología*, (10), 12-20. <https://doi.org/10.17163/ings.n10.2013.02>
- Microsoft Corporation. (2025). *Introducción a Speaker Recognition - Azure AI services* [Consultado el 17 de agosto de 2025. [Documentación modificada o descontinuada por el proveedor]]. Consultado el 17 de agosto de 2025, desde <https://learn.microsoft.com/es-es/azure/ai-services/speechservice/speaker-recognition-overview#speaker-identification>
- Mohamed, A. A., Eltokhy, A., & Zekry, A. A. (2022). An Overview of the Development of Speaker Recognition Techniques for Various Applications. *Journal of Communications*, 17(8), 632-642. <https://doi.org/10.12720/jcm.17.8.632-642>
- Moreno & Rüdinger. (2024). Avances y desafíos en el reconocimiento de voz en alemán: aplicaciones y contrastes. En *Exploraciones en el Idioma, Literatura y Traducción del Alemán: Desde Plataformas de Aprendizaje hasta Expresiones Culturales* (p. 27). Centro Universitario EUSA.
- Nagawade, M. S., & Ratnaparkhe, V. R. (2017). Musical instrument identification using MFCC. *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 2198-2202. <https://doi.org/10.1109/RTEICT.2017.8256990>
- Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A Large-Scale Speaker Identification Dataset. *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2616-2620. <https://doi.org/10.21437/Interspeech.2017-950>

- Nouhaila, B., Taoufiq, B. D., & Benayad, N. (2022). An Intelligent Approach based on the Combination of the Discrete Wavelet Transform, Delta Delta MFCC for Parkinson's Disease Diagnosis. *International Journal of Advanced Computer Science and Applications*, 13(4). <https://doi.org/10.14569/IJACSA.2022.0130466>
- Olguín Meza, M. d. J. (2023). ¿Cómo viaja el sonido? [Material didáctico sobre cualidades del sonido: intensidad, tono y timbre]. <https://dspace.uaeh.edu.mx/server/api/core/bitstreams/b16e9972-8a35-4f2c-8aca-679351220986/content>
- Oppenheim, A. V., & Schaffer, R. W. (2011). *Tratamiento de señales en tiempo discreto* (3.^a ed.). Pearson Educación.
- Ortega-Mendoza, R. M., Villatoro-Tello, E., Villaseñor-Pineda, L., Montes-y-Gómez, M., & Morales, E. F. (2007). On the Use of Dynamic Information for Speaker Identification [Trabajo realizado con el corpus DIMEx100, apoyado por CONACYT (Proyecto 43990). Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).]. *CORE 2007 Workshop*.
- Palacios Sáez-Guinea, L. (2022, junio). *Aprendizaje automático y reglas de asociación* [Trabajo de fin de grado de Matemáticas]. Universidad de Zaragoza [Director del trabajo: Ricardo López Ruiz].
- Pineda, L. A., Pineda, L. V., Cuétara, J., Castellanos, H., & López, I. (2004). DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish. *Advances in Artificial Intelligence – IBERAMIA 2004*, 974-983. <https://doi.org/10.1007/978-3-540-30498-2>
- Ramírez, J., & Ramos, M. (2015). Clasificación mediante máquinas de soporte vectorial. *Matemática*, 13(2), 29-40.
- Real Academia Española. (2026). *Diccionario de la lengua española* (23.^a ed., Ver. Versión 23.8.1 en línea) [Consultado para la definición de frecuencia]. <https://dle.rae.es>
- Sandoval Serrano, L. J., et al. (2018). Algoritmos de aprendizaje automático para análisis y predicción de datos. *Revista Tecnológica; no. 11*.
- Shaveta. (2023). A Review on Machine Learning. *International Journal of Science and Research Archive*, 9(1), 281-285. <https://doi.org/10.30574/ijrsra.2023.9.1.0410>

- Sintec Insonorización. (2021). *Conceptos fundamentales del sonido y bandas frecuenciales* [Manual técnico de aislamiento acústico y bandas de frecuencia]. Sintec Insonorización. Barcelona, España. <https://www.acdacustics.com/files/conceptos.pdf>
- Spotfire Community. (s.f.). *Random forest template for Spotfire* [[En línea; accedido el 22-junio-2026]]. <https://community.spotfire.com/articles/spotfire/random-forest-template-for-spotfire/>
- Stevens, S., Volkman, J., & Newman, E. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 8(3), 185-190.
- Stryker, C., & Holdsworth, J. (2024, agosto). *¿Qué es el procesamiento de lenguaje natural (PLN)?* [Consultado el 24 de agosto de 2025]. IBM. <https://www.ibm.com/mx-es/think/topics/natural-language-processing>
- Syed, I., & Lokhande, V. (2024). An Overview of the Supervised Machine Learning. *International Research Journal of Modernization in Engineering Technology and Science*, 6(3), 6355-6360. <https://doi.org/10.56726/IRJMETTS51366>
- The Voice Foundation. (2019, febrero). *Understanding voice production*. <https://voicefoundation.org/health-science/voice-disorders/anatomy-physiology-of-voice-production/understanding-voice-production/>
- Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*, 90, 250-271. <https://doi.org/10.1016/j.eswa.2017.08.015>
- Vázquez Robledo, A. S., Lizárraga Morales, R. A., & López Ramírez, M. (2024). Clasificación de instrumentos musicales en audios utilizando coeficientes cepstrales y redes neuronales artificiales. *Revista Jóvenes en la Ciencia*. <http://repositorio.ugto.mx/handle/20.500.12059/14110>
- Wirdiani, A., Ndung'u, S., Putra, I., Sudarma, M., Hartati, R., & Ferdian, H. (2024). Improvement Model for Speaker Recognition using MFCC-CNN and Online Triplet Mining. *Internatio-*

- nal Journal on Advanced Science, Engineering and Information Technology*, 14, 420-427.
<https://doi.org/10.18517/ijaseit.14.2.19396>
- Yan, Y., Simons, S. O., van Bommel, L., Reinders, L., Franssen, F. M. E., & Urovi, V. (2024). Optimising MFCC parameters for the automatic detection of respiratory diseases. <https://arxiv.org/abs/2408.07522>
- Youngsik, E., & Bang, J. (2021). Speech Emotion Recognition Using 2D-CNN with Mel-Frequency Cepstrum Coefficients. *Journal of Information and Communication Convergence Engineering*, 19(3), 148-154. <https://doi.org/10.6109/jicce.2021.19.3.148>
- Zhang, Y., & Liu, L. (2023). Multi-task learning for X-vector based speaker recognition. *International Journal of Speech Technology*, 26(4), 817-823. <https://doi.org/10.1007/s10772-023-10058-5>
- Zheng, S., & Ding, C. (2018). Minimal Support Vector Machine. <https://arxiv.org/abs/1804.02370>
- Zhu, W., & Han, S. (2022). Cultural Product Appearance Design Based on Improved Multiobjective Optimization Algorithm. *Security and Communication Networks*, 2022, 1-9. <https://doi.org/10.1155/2022/7063303>