

UNIVERSIDAD AUTÓNOMA DE ZACATECAS
“Francisco García Salinas”



**“Aplicación de técnicas de filtrado y adaptación de dominio en la señal de voz
con fines de reconocimiento del habla en entornos con ruido”**

Tesis para obtener el grado de:
Maestro en Ciencias del Procesamiento de la Información

Presenta

I.C. Emmanuel De Jesús Velásquez Martínez

Director:

Dr. Efrén González Ramírez

Co-Directores:

Dr. Aldonso Becerra Sánchez

Dr. José Ismael De La Rosa Vargas

Asesores:

Dr. Gamaliel Moreno Chávez

Dr. Daniel Alaniz Lumbreras

Zacatecas, Zac., 27 de octubre de 2023



SOMOS
ARTE, CIENCIA Y
DESARROLLO
CULTURAL



Zacatecas, Zac., 26 de octubre de 2023.

C. Emmanuel De Jesús Velásquez Martínez
Estudiante de la MCPI
PRESENTE

At'n: Dr. Huizilopoztli Luna García
Responsable de la MCPI

Nos es grato comunicarle que después de haber sometido a revisión académica la propuesta de Tesis titulada "**Aplicación de Técnicas de Filtrado y Adaptación de Dominio en la Señal de Voz con Fines de Reconocimiento del Habla en Entornos con Ruido**", presentada por el estudiante Ing. Emmanuel De Jesús Velásquez Martínez y habiendo efectuado todas las correcciones indicadas por este Comité Tutorial, se **AUTORIZA** el documento de tesis para su impresión.

Sin más por el momento reciban un cordial saludo.

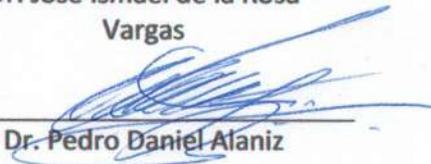
COMITÉ TUTORIAL
PROCESAMIENTO Y ANÁLISIS DE DATOS


Dr. Efrén González Ramírez


Dr. José Ismael de la Rosa
Vargas


Dr. Aldonso Becerra Sánchez


Dr. Gamaliel Moreno
Chávez


Dr. Pedro Daniel Alaniz
Lumbreras

c.c.p. Interesado.

c.c.p. Responsable de la Maestría en Ciencias del Procesamiento de la Información.



SOMOS
ARTE, CIENCIA Y
DESARROLLO
CULTURAL

**COORDINACIÓN DE
INVESTIGACIÓN Y POSGRADO**

Carta de similitud núm. 460/ IyP
Zacatecas, Zacatecas 17/octubre/2023

Dr. Huizilopoztli Luna García
Responsable de la MCPI – UAZ
Presente

Estimado Dr. Huizilopoztli,

Después de saludarlo, sirva el presente oficio para notificar que el documento

*"Aplicación de técnicas de filtrado y adaptación de dominio en la señal de voz
con fines de reconocimiento del habla en entornos con ruido"*
De Emmanuel De Jesús Velásquez Martínez

Fue analizado con el software Copyleaks, con la intención de detectar similitudes; el resultado en cuestión fue

7.9 % de similitud

De acuerdo a lo anterior, el porcentaje se considera **ACEPTABLE** de acuerdo a los estándares internacionales.

Atentamente

"Somos Arte, Ciencia y Desarrollo Cultural"

Dr. Carlos Francisco Bautista Capetillo
Coordinador de Investigación y Posgrado
Universidad Autónoma de Zacatecas



SOMOS
ARTE, CIENCIA Y
DESARROLLO
CULTURAL



Zacatecas, Zac., 26 de octubre de 2023

Carta de cesión de derechos

A QUIEN CORRESPONDA

El suscrito, C. Emmanuel de Jesús Velásquez Martínez, alumno del Programa de Maestría en Ciencias del Procesamiento de la Información, con número de matrícula 35161479 y adscrito a la Unidad Académica de Ingeniería de la Universidad Autónoma de Zacatecas, desea informar que soy el autor intelectual del trabajo de Tesis titulado "Aplicación de técnicas de filtrado y adaptación de dominio en la señal de voz con el propósito de reconocimiento del habla en entornos con ruido". Bajo la dirección del Dr. Efrén González Ramírez, cedo los derechos de este trabajo a la Universidad Autónoma de Zacatecas con el propósito de su difusión para fines académicos e investigativos.

Se solicita a los usuarios de esta información que se abstengan de reproducir el contenido textual, gráficos o datos del trabajo sin la autorización expresa del autor y/o los directores de la investigación. Para obtener dicho permiso, pueden contactarme a través del correo electrónico iemmanuelvm@gmail.com o comunicarse con el responsable del Programa de Maestría, quien derivará la solicitud a los directores del trabajo de investigación. Si se concede el permiso, se espera que el usuario muestre la debida gratitud y cite la fuente apropiadamente.

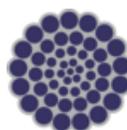
Agradezco de antemano su atención a esta solicitud y quedo a su disposición para cualquier consulta adicional. Reciba un cordial saludo.

ATENTAMENTE

Emmanuel de Jesús Velásquez Martínez



SOMOS
ARTE, CIENCIA Y
DESARROLLO
CULTURAL



CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES
CIENCIAS Y TECNOLOGÍAS

AGRADECIMIENTO ESPECIAL

Al Programa Nacional de Posgrados de Calidad del Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) por su apoyo económico a través de la convocatoria Nacional (Tradicional) 2021 2023.

Resumen

El reconocimiento de voz en la actualidad es una tarea muy común en diversos sistemas cotidianos de usuario, sin embargo, carece de buena efectividad en entornos con ruido, como autos en movimiento, hogares con ruido ambiental, teléfonos móviles, entre otros. Este trabajo combina técnicas de aprendizaje profundo con adaptación de dominio y filtrado basado en la transformada Wavelet para eliminar el ruido estacionario y no estacionario en las señales de voz. El enfoque empleado tiene como objetivo abordar el reconocimiento automático de voz (RAV) y la identificación de locutor en entornos ruidosos. Este trabajo demuestra cómo un modelo de redes neuronales profundas con adaptación de dominio puede mitigar diversos tipos de ruido. Una de las teorías a aplicar es el uso del Transporte Óptimo tanto en la tarea de regresión para mejora del habla ruidosa y para la tarea de identificación de locutor; es así como que la aplicación de esta teoría en aprendizaje profundo ha demostrado mejorar la eficiencia para entrenar un modelo de aprendizaje profundo. Las evaluaciones del habla se realizaron con base a la inteligibilidad objetiva a corto plazo (STOI) y calidad de la evaluación perceptual del habla (PESQ). Se aplicó la transformada wavelet (TW) como técnica de filtrado para realizar un segundo procesamiento en la señal mejorada por la red neuronal profunda, que alcanzó en promedio una mejora del 20% en STOI y un 9% en PESQ respecto a la señal ruidosa. Por último, se evaluó el método en un esquema de RAV preentrenado, logrando una disminución general de la tasa de error de palabra a 14.24% y alcanzando en promedio un 99% en la identificación de locutor. El enfoque propuesto proporciona una mejora significativa en el rendimiento del reconocimiento del habla al abordar el problema del habla ruidosa en diversos entornos.

Abstract

Speech recognition today is a very common task in various everyday user systems; however, it lacks good effectiveness in noisy environments, such as moving vehicles, homes with background noise, mobile phones, among others. This work combines deep learning techniques with domain adaptation and Wavelet transform-based filtering to remove both stationary and non-stationary noise from speech signals. The approach employed aims to address automatic speech recognition (ASR) and speaker identification in noisy environments. This work demonstrates how a deep neural network model with domain adaptation can mitigate various types of noise. One of the theories to apply is the use of Optimal Transport in both the speech enhancement regression task and the speaker identification task; thus, the application of this theory in deep learning has been shown to improve the efficiency of training a deep learning model. Speech evaluations were conducted based on Short-Time Objective Intelligibility (STOI) and Perceptual Evaluation of Speech Quality (PESQ). The Wavelet transform (WT) was applied as a filtering technique to perform a second processing on the speech signal enhanced by the deep neural network, which on average achieved a 20% improvement in STOI and a 9% improvement in PESQ compared to the noisy signal. Finally, the method was evaluated in a pre-trained ASR scheme, achieving an overall word error rate reduction to 14.24% and an average of 99% in speaker identification. The proposed approach provides a significant improvement in speech recognition performance by addressing the problem of noisy speech in various environments.

Contenido general

Capítulo 1	Introducción	1
1.1	Antecedentes	1
1.2	Planteamiento del problema.....	2
1.3	Justificación del problema de investigación	3
1.4	Preguntas de investigación.....	4
1.5	Objetivo general.....	4
1.6	Objetivos específicos	4
1.7	Hipótesis	4
1.8	Alcance	5
1.9	Estructura de la tesis	5
Capítulo 2	Reconocimiento del habla.....	6
2.1	Descripción de los sistemas del reconocimiento del habla	6
2.2	Principales desafíos de los reconocedores del habla.....	6
2.3	Métodos para el reconocimiento del habla.....	7
2.3.1	Reconocimiento de patrones	8
2.3.2	Sistemas basados en el conocimiento	8
2.3.3	Modelos estocásticos.....	8
2.3.4	Modelos neuronales o conexionistas.....	8
2.4	Descripción de las teorías base de un sistema de reconocimiento de voz.....	8
2.4.1	Formulación matemática.....	8
2.4.2	Procesamiento acústico	9
2.4.3	Modelado acústico	9
2.4.4	Modelo de lenguaje.....	10
2.4.5	Hipótesis de búsqueda.....	10
2.5	Esquemas de reconocimiento de voz	10
2.5.1	Modelos híbridos - Mezclas Gaussianas – Modelos Ocultos de Markov	11
2.5.2	Redes Neuronales Artificiales – Modelos Ocultos de Markov	11
2.5.3	Modelos End-to-End.....	12
2.6	Descripción de las teorías base de un sistema para identificación de locutor	13
2.6.1	Identificación de locutor	13
2.6.2	Procesamiento acústico	14
2.6.3	Extracción de características	14
2.6.4	Modelos para identificación de locutor.....	18
Capítulo 3	Adaptación de la señal del habla para reducción de ruido	19
3.1	Ruido.....	19

3.1.1	Tipos de ruido	19
3.2	Adaptación de dominio	19
3.2.1	Introducción a la teoría de Transporte Óptimo para la adaptación de dominio	20
3.2.2	Tipos de Transporte Óptimo	21
3.2.3	Formulaciones del Transporte Óptimo.....	21
3.2.4	Transporte Óptimo en redes neuronales profundas.....	23
3.2.5	Adaptación de dominio con Transporte Óptimo en modelos de aprendizaje profundo	24
3.3	Modelos de aprendizaje profundo para regresión y clasificación.....	26
3.3.1	Redes adversarias generativas.....	26
3.3.2	Red adversaria generativa como modelo de regresión.....	27
3.3.3	Red adversaria generativa como clasificador.....	28
3.3.4	Descripción de las arquitecturas de redes neuronales en las GANs.....	28
3.4	Trasformada Wavelet.....	30
3.4.1	Definición de Transformada Discreta Wavelet y sus propiedades	30
3.4.2	Transformada Wavelet Discreta.....	31
3.4.3	Algoritmo para aplicar la técnica de filtrado basado en la transformada Wavelet.....	32
3.4.4	Niveles de descomposición.....	33
3.4.5	Umbralización.....	34
3.5	Estudios relacionados.....	35
3.5.1	Estudios relacionados sobre los reconocedores de voz.....	35
3.5.2	Estudios relacionados sobre la identificación de locutor	35
3.5.3	Estudios relacionados de adaptación de dominio en el reconocimiento voz	37
3.5.4	Estudios relacionados sobre la mejora del habla con transformada Wavelet.....	39
3.5.5	Estudios relacionados sobre la mejora del habla con aprendizaje profundo.....	39
Capítulo 4	Método y propuesta de investigación.....	41
4.1	Modelo de investigación	41
4.2	Propuesta de trabajo a realizar	42
4.2.1	Descripción de la propuesta de trabajo	42
4.3	Selección y preparación de los conjuntos de datos	43
4.3.1	Corpus para el reconocimiento del habla	43
4.3.2	Corpus con ruido natural.....	43
4.3.3	Generación del conjunto de habla ruidosa	44
4.4	Configuración del problema del habla ruidosa	45
4.4.1	Configuración computacional para extraer los coeficientes de la transformada de Fourier en tiempo corto	46
4.5	Configuración del problema de adaptación de dominio	47
4.5.1	Adaptación de dominio como problema de regresión.....	47

4.5.2	Configuración computacional para la implementación de la red neuronal de mejora de la señal de voz.....	48
4.6	Evaluación de la señal de voz	50
4.6.1	STOI.....	50
4.6.2	PESQ.....	50
4.7	Configuración de la transformada Wavelet como técnica de filtrado para eliminación de ruido51	
4.8	Modelos del reconocimiento del habla	51
4.8.1	Configuración del modelo de identificación de locutor	51
4.8.2	Configuración del modelo reconocedor automático de voz.....	59
4.9	Hardware y software utilizado	59
4.9.1	Recursos de hardware	59
4.9.2	Recursos de software	60
Capítulo 5	Resultados y limitaciones.....	61
5.1	Comportamiento de la función de pérdida en los dominios fuente y objetivo.....	61
5.2	Resultados de eliminación de ruido en la señal de voz del dominio fuente.....	65
5.2.1	Resultados para tipo de ruido estacionario – ruido rosa.....	65
5.2.2	Resultados para tipo de ruido estacionario – gotas de agua.....	66
5.2.3	Resultados para tipo de ruido estacionario – carro	67
5.2.4	Resultados para tipo de ruido estacionario – cabina	68
5.2.5	Resultados para tipo de ruido estacionario – lluvia	69
5.2.6	Resultados para tipo de ruido estacionario – viento.....	71
5.2.7	Resultados para tipo de ruido estacionario – escritura de teclado.....	72
5.3	Resultados de eliminación de ruido en la señal de voz del dominio objetivo.....	74
5.3.1	Resultados para tipo de ruido no estacionario – llanto de bebé	74
5.3.2	Resultados para tipo de ruido no estacionario – fiesta con multitud de gente	75
5.3.3	Resultados para tipo de ruido no estacionario – campanas de iglesia.....	76
5.3.4	Resultados para tipo de ruido no estacionario – mormullos en cafetería.....	77
5.3.5	Resultados para tipo de ruido no estacionario – helicóptero.....	79
5.3.6	Resultados para tipo de ruido no estacionario – personas hablando	80
5.3.7	Resultados para tipo de ruido no estacionario – ladrido de perro	81
5.4	Reconocimiento automático de voz	82
5.4.1	Resultados para datos fuente.....	82
5.4.2	Resultados para datos objetivo.....	90
5.5	Identificación de locutor	97
Capítulo 6	Conclusiones	101
6.1	Objetivos alcanzados	101
6.2	Hipótesis/proposiciones demostradas	102

6.3	Contribuciones de la investigación	103
6.4	Trabajos futuros	103
	Referencias.....	104
	Anexos	110

Índice de figuras

Figura 2.1 Arquitectura de un sistema reconocedor de voz	6
Figura 2.2 Los problemas actuales de un RAV (columna derecha) son mucho más difíciles que en los que hemos trabajado en el pasado debido a la demanda de las aplicaciones del mundo real	7
Figura 2.3 Arquitectura de un sistema de reconocimiento automático de voz MMG-MOM	12
Figura 2.4 Arquitectura de un sistema de reconocimiento automático de voz RNP-MOM	12
Figura 2.5 Arquitectura de un sistema de reconocimiento automático de voz End-to-End	13
Figura 2.6 Extracción de MFCCs	15
Figura 3.1 La imagen de la izquierda muestra dos medidas μ y ν en \mathbf{X} dadas como densidades. La derecha muestra un subconjunto medible $\mathbf{A} \subseteq \mathbf{X}$ y su imagen inversa bajo un mapa de transporte \mathbf{T}	20
Figura 3.2 Red neuronal adversaria generativa.....	27
Figura 3.3 Red neuronal adversaria generativa como modelo de aprendizaje de regresión	28
Figura 3.4 Red neuronal adversaria generativa - EC-GAN	28
Figura 3.5 Red neuronal convolucional	29
Figura 3.6 Arquitectura de red neuronal convolucional VGG-16.....	30
Figura 3.7 Red neuronal recurrente bidireccional.....	31
Figura 3.8 Ejemplos de Wavelets	33
Figura 3.9 Niveles de descomposición de la Transformada Wavelet	34
Figura 3.10 Tipos de umbralización	34
Figura 4.1 Modelo de investigación general.....	41
Figura 4.2 Propuesta de trabajo para el sistema de reconocimiento del habla en entornos ruidosos.....	42
Figura 4.3 Generación del conjunto del habla ruidosa.....	44
Figura 4.4 Arquitectura de red neuronal profunda con adaptación de dominio para mejora del habla ruidosa	48
Figura 4.5 Técnica de filtrado basado en transformada Wavelet.....	51
Figura 4.6 Modelos del reconocimiento del habla.....	52
Figura 4.7 Arquitectura de red neuronal profunda para clasificación de locutor.....	53

Figura 5.1 Comportamiento de función de pérdida ECM para mejora del habla en el dominio origen	61
Figura 5.2 Comportamiento de función de pérdida ECM para mejora del habla en el dominio destino	62
Figura 5.3 Formas de onda de la señal de voz en el dominio del tiempo para un enunciado pronunciado por un locutor del conjunto de datos LibriSpeech en su versión base. Esta señal está contaminada con el ruido de helicóptero proporcionado en un SNR nivel -3 dB y sus versiones mejoradas.....	63
Figura 5.4 Comportamiento del espectro de la señal de voz para un enunciado pronunciado por un locutor de LibreSpeech en su versión base. Esta señal está contaminada con el ruido de helicóptero proporcionado en un SNR nivel -3 dB y sus versiones mejoradas.....	64
Figura 5.5 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: rosa	65
Figura 5.6 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: gotas de agua	66
Figura 5.7 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: carro	68
Figura 5.8 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: cabina	69
Figura 5.9 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: lluvia.....	70
Figura 5.10 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: viento.....	72
Figura 5.11 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: escritura de teclado.....	73
Figura 5.12 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: llanto de bebe	74
Figura 5.13 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: fiesta de multitud de gente	76
Figura 5.14 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: campanas de iglesia.....	77
Figura 5.15 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: mormullos de cafetería... ..	78
Figura 5.16 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: helicóptero.....	79
Figura 5.17 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: personas hablando	80
Figura 5.18 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: ladrido de perro	81
Figura 5.19 Comportamiento de la tasa de error de palabra - tipo de ruido: rosa.....	83
Figura 5.20 Comportamiento de la tasa de error de palabra - tipo de ruido: gotas de agua.....	84
Figura 5.21 Comportamiento de la tasa de error de palabra - tipo de ruido: carro	85
Figura 5.22 Comportamiento de la tasa de error de palabra - tipo de ruido: cabina	86
Figura 5.23 Comportamiento de la tasa de error de palabra - tipo de ruido: lluvia	87
Figura 5.24 Comportamiento de la tasa de error de palabra - tipo de ruido: viento.....	88

Figura 5.25 Comportamiento de la tasa de error de palabra - tipo de ruido: escritura de teclado	89
Figura 5.26 Comportamiento de la tasa de error de palabra - tipo de ruido: llanto de bebe	90
Figura 5.27 Comportamiento de la tasa de error de palabra - tipo de ruido: multitud de gente.....	91
Figura 5.28 Comportamiento de la tasa de error de palabra - tipo de ruido: campanas de iglesia.....	92
Figura 5.29 Comportamiento de la tasa de error de palabra - tipo de ruido: mormullos de cafetería.....	93
Figura 5.30 Comportamiento de la tasa de error de palabra - tipo de ruido: helicóptero.....	94
Figura 5.31 Comportamiento de la tasa de error de palabra - tipo de ruido: personas hablando	95
Figura 5.32 Comportamiento de la tasa de error de palabra - tipo de ruido: ladrido de perro	96
Figura 5.33 Comparación del comportamiento de la función de perdida de entropía cruzada durante el entrenamiento de la red neuronal	98
Figura 5.34 Comparación del comportamiento de la métrica de exactitud durante el entrenamiento de la red neuronal	99
Figura 5.35 Matrices de confusión para cada conjunto de datos de prueba.....	100

Índice de tablas

Tabla 4.1 Conjuntos de ruido	44
Tabla 4.2 Arquitectura del generador para eliminación de ruido.....	49
Tabla 4.3 Arquitectura del discriminador para eliminación de ruido	49
Tabla 4.4 Métrica STOI	50
Tabla 4.5 Métrica PESQ	50
Tabla 4.6 Arquitectura del generador	54
Tabla 4.7 Arquitectura del discriminador	55
Tabla 4.8 Arquitectura de red neuronal para clasificador de locutores.....	56
Tabla 4.9 Especificaciones del equipo	60
Tabla 4.10 Software empleado para la experimentación	60
Tabla 5.1 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: rosa	66
Tabla 5.2 Puntajes STOI y PESQ para corpus LibreSpeech - Tipo de ruido: Gotas de agua.....	67
Tabla 5.3 Puntajes STOI y PESQ para corpus LibreSpeech - Tipo de ruido: Carro.....	68
Tabla 5.4 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: cabina.....	70
Tabla 5.5 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: lluvia.....	71
Tabla 5.6 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: viento	72
Tabla 5.7 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: escritura de teclado	73
Tabla 5.8 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: llanto de bebe.....	75
Tabla 5.9 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: fiesta de multitud de gente... 76	
Tabla 5.10 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: campanas de iglesia	77
Tabla 5.11 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: mormullos de cafetería	79
Tabla 5.12 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: helicóptero	80
Tabla 5.13 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: personas hablando	81
Tabla 5.14 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: ladrido de perro	82
Tabla 5.15 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: rosa.....	83

Tabla 5.16 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: gotas de agua.....	84
Tabla 5.17 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: carro	85
Tabla 5.18 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: cabina	86
Tabla 5.19 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: lluvia	87
Tabla 5.20 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: viento	88
Tabla 5.21 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: escritura de teclado	89
Tabla 5.22 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: llanto de bebe	91
Tabla 5.23 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: multitud de gente.....	92
Tabla 5.24 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: campanas de iglesia.....	93
Tabla 5.25 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: mormullos de cafetería.....	94
Tabla 5.26 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: helicóptero.....	95
Tabla 5.27 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: personas hablando	96
Tabla 5.28 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: ladridos de perro	97
Tabla 5.29 Métricas de evaluación para el modelo de identificación de locutor	99

Capítulo 1 Introducción

Este primer capítulo consta de nueve partes centrales que fijan los fundamentos del estudio. Se presentan los antecedentes, la justificación del problema de investigación, preguntas de investigación, objetivo general, objetivos específicos, hipótesis y la estructura de la tesis.

1.1 Antecedentes

La tecnología del reconocimiento del habla, se puede dividir en dos subáreas principales, que son el reconocimiento de voz, que analiza el contenido de palabras y la identificación del locutor, que asemeja quién es el hablante de dichas palabras (Bunrit et al., 2019). El cual de acuerdo con Yu & Deng, (2014), ha sido un área de investigación activa durante más de cinco décadas, donde siempre se ha considerado como un puente importante para fomentar una mejor comunicación entre humanos-humanos y entre humanos-máquinas. En los últimos años, la tecnología del habla comenzó a cambiar la forma en que vivimos y trabajamos, y se convirtió en uno de los principales medios para que los humanos interactúen con algunos dispositivos. Partiendo de lo anterior, los sistemas de reconocimientos del habla implican el proceso capaz de convertir la señal de voz en una secuencia de palabras o comandos, en donde bajo la implementación de algoritmos se realiza un proceso de reconocimiento de voz para llevar a cabo acciones informáticas (Mohamed et al., 2020). Se incorporan técnicas de procesamiento de señales y aprendizaje automático para reconocer el habla. Sin embargo, los sistemas tradicionales tienen un bajo rendimiento debido a los factores que no se tomaron en cuenta, y que por consiguiente afectan negativamente el rendimiento del sistema de reconocimiento del habla al analizar las señales de voz (Ali et al., 2022).

Existen una gran variedad de aplicaciones del reconocimiento de voz automático (RAV) tales como atención al cliente, respuesta de voz interactiva, ingreso y dictado de datos, comando y control mediante voz, creación de documentos estructurados, control de dispositivos mediante voz, aprendizaje de idiomas, contenido de búsqueda basado en audio hablado, robótica, internet de las cosas y sistemas biométricos, entre otros. Estas aplicaciones de RAV son susceptibles a diferentes factores e influyen dentro de estos sistemas para obtener un rendimiento deseado de reconocimiento del habla; entre estos factores tenemos al ruido (Jinyu et al., 2014; Nossier et al., 2021; Roy et al., 2021). Las señales ruidosas se generan por los diversos entornos y se pueden mezclar por naturaleza a la señal de voz. Lo anterior ocasiona que la interferencia de ruido disminuya severamente en calidad perceptiva e inteligibilidad en las tareas de reconocimiento del habla (Adeel et al., 2020). Es por ello que, para aplicaciones del mundo real, la robustez del ruido se ha convertido en una tecnología de interés cada vez más importante, ya que los sistemas de reconocimiento del habla necesitan funcionar en entornos acústicos mucho más influenciados por el ruido (Vikramjit et al., 2011).

A consecuencia del factor de ruido en los sistemas de RAV, se han estado dedicando esfuerzos para mejorar el robustecimiento y aumentar las tasas de reconocimiento del habla. Existen principalmente dos enfoques para la robustez del ruido. El primer enfoque consiste en métodos capaces de mejorar las características extraídas del audio, el objetivo es eliminar el ruido de las observaciones antes del reconocimiento de voz (Vazhenina & Markov, 2020). Por otro lado, se tienen los enfoques de adaptación del modelo donde no se modifican las observaciones de la señal de voz, si no que se tratan de mejorar. El objetivo de este enfoque se basa en que los parámetros

de la señal del habla sean más representativos y que se asemejen a la señal de voz original. Los estudios de este tipo de mejora de la señal de voz tienen como objetivo la mejora del habla en entornos ruidosos en cuanto a la calidad y la intangibilidad del habla corrupta por ruido (Choi et al., 2019; Leglaive et al., 2019). De ahí que es esencial buscar, aplicar y/o combinar técnicas efectivas para mejorar la calidad de la señal, para que en ese sentido se aumente la precisión de los sistemas de reconocimiento de habla en diferentes entornos ruidosos. Actualmente, los modelos de aprendizaje profundo y técnicas de filtrado son aplicables para este tipo de escenarios (Adeel et al., 2020); sin embargo, el desafío que se tiene actualmente dados los diferentes entornos acústicos, radica en que la implementación de estos modelos se puede comportar de forma diferente respecto a la experimentación, y los ruidos no previstos pueden degradar seriamente la calidad de la señal procesada (Liao et al., 2019). Este desajuste generalmente se denomina problema de desajuste de dominio. Entonces se requiere llevar a cabo adaptación del dominio en redes neuronales para tener un modelo de aprendizaje capaz de adaptarse a diferentes tipos de ruidos (Lin et al., 2021).

La implementación de este trabajo tiene como objetivo estimar las referencias limpias de voz con ruido en dominios distintos. Demostrando tener en los resultados experimentales una mejora en el procesamiento de voz, superando a otros métodos que se encuentran en el estado del arte. Finalmente es importante mencionar que, en la literatura, el RAV tiene una línea base de rendimiento en tasa de error de palabra del 10 al 25% en entornos ruidosos no controlados. Como antecedentes de lo anterior, en el presente trabajo se propone un enfoque de adaptación de dominio para tratar la discrepancia entre el conjunto de entrenamiento y el de pruebas, tomando como base una señal de voz ruidosa, aplicando además filtros para eliminación de ruido.

1.2 Planteamiento del problema

El problema del reconocimiento de voz en condiciones ruidosas es muy difícil y diverso. Este constituye un importante cuello de botella para el uso práctico de los reconocedores de voz en condiciones reales, ya que hasta el momento no se ha propuesto una solución completamente satisfactoria. Las dificultades del reconocimiento del habla ruidosa, provienen de los diversos efectos del ruido en el habla que se pueden resumir como la adición de ruido ambiental a la señal del habla, distorsión de la señal y variaciones en la articulación. Esto genera como resultado dos fenómenos principales: la degradación de las actuaciones de un sistema de reconocimiento de voz y las condiciones para el aprendizaje (Haton, 1994). Por esa razón los retos del reconocimiento de voz en la actualidad, debido a la gran cantidad de aplicaciones en donde se involucran factores que afectan en la tarea de reconocimiento de voz, tales como la distorsión y el ruido del entorno han sido de investigación activa. Se han propuesto un gran número de métodos para hacer frente a este problema, aunque ninguno es totalmente satisfactorio, además cada uno presenta ventajas y desventajas (Liu et al., 2019). Las técnicas de filtrado han sido una herramienta muy útil para eliminar el ruido de una amplia variedad de señales. Pero la mejora de la señal a través de técnicas de filtrado solo mejora hasta cierto nivel la señal de voz deseada, eso hace que sea necesario aumentar la solidez general del reconocimiento de voz.

La solución simple consistiría en tener las mismas condiciones para el entrenamiento y la prueba en cuanto a las señales limpias de ruido, pero rara vez es asequible y, por lo general, no es realista; la modificación de sonidos pronunciados en un ambiente ruidoso. A pesar de los resultados significativos, el problema aún no está completamente resuelto mediante estas técnicas, si no aún queda por hacer un esfuerzo importante para evaluar comparativamente los diferentes

métodos existentes y mejorarlos. Este efecto depende en gran medida del hablante y el nivel de ruido. Por otro lado, el modelo de aprendizaje para el reconocimiento de voz automático en el que influye el ruido puede afectar al momento en el que se implementa en un entorno real en el que no se estuvieron contemplados esos factores o en el cual difieren por las condiciones en que fue entrenado el modelo de aprendizaje (Fan et al., 2020). Por lo tanto, es difícil modelizar, y la falta de coincidencias entre los patrones de entrenamiento y prueba es otro de los principales problemas que se debe enfrentar en condiciones adversas de la vida real del reconocimiento de voz. Debido a esto, la falta de coincidencias de las distribuciones estadísticas obtenidas de la voz entre los conjuntos de entrenamiento y prueba, el rendimiento del reconocimiento automático de voz tiende a degradarse (J. Li et al., 2014).

Es por ello que el principal reto es implementar técnicas de filtrado como parte de un preprocesamiento de la señal de voz y adaptación de dominio en redes neuronales profundas, el cual ayude a mejorar el esquema de reconocimiento de voz en entornos con ruido. Donde al minimizar el ruido y la pérdida de clasificación del modelo de aprendizaje se adapte en diversos entornos, obteniendo un mayor rendimiento en la tasa de error de palabra de un RAV.

1.3 Justificación del problema de investigación

La implementación de esta propuesta se justifica debido a la amplia variedad de aplicaciones que involucran el habla en la vida cotidiana de los usuarios. El ruido ambiental está presente en cualquier entorno y puede mezclarse fácilmente con la señal de voz, lo que resulta en una degradación en la tasa de reconocimiento. Por lo tanto, la eliminación de ruido del habla puede mejorar significativamente la percepción en un sistema de reconocimiento automático del habla, proporcionando un contexto más preciso en la comunicación de la información. Esto permite realizar acciones más precisas mediante el uso de la voz. La propuesta de acoplar dos técnicas para el tratamiento del ruido en este trabajo tiene como objetivo principal reducir el ruido mientras se recuperan las señales originales. Aunque existen varias técnicas propuestas para eliminar el ruido de una señal de audio, la eficiencia sigue siendo un desafío en la mayoría de ellas. Por lo tanto, se espera lograr mejoras sustanciales en la robustez de los sistemas de reconocimiento de voz al combinar estas dos técnicas de reducción de ruido.

Cuando se desarrolla un algoritmo capaz de identificar el lenguaje hablado tiende a tener valor dentro de sus aplicaciones de uso. Este tipo de algoritmos desarrollados asumen que hay un conjunto de datos a los que se denominan datos de entrenamiento y de prueba, el cual comparten una propiedad de distribución similar. Pero debido a que existen distintos patrones tanto acústicos como lingüísticos, esto ocasiona que los conjuntos tanto de entrenamiento y prueba sean distintos entre sí, además debido a la adición de ruido estos tienden a ser diferentes. Por lo tanto, un método es acoplar la distribución del conjunto de datos de prueba para que de esta forma se tenga una precisión respecto a los datos de entrenamiento. Esto con el fin de tener mejoras en el reconocimiento del habla. Dado que este tipo de aprendizaje es fundamental para los avances futuros, se busca semigeneralizar o generalizar los modelos de aprendizaje automático para obtener un mejor rendimiento en el reconocimiento del habla.

La implementación de esta propuesta tiene como objetivo abordar los desafíos mencionados y contribuir al desarrollo de sistemas de reconocimiento de voz más efectivos y robustos. Por lo tanto, el objetivo de esta justificación es aumentar la eficiencia de absorción de diferentes ruidos respecto a los niveles de SNR (relación señal ruido), para reducir el riesgo de reconocimiento de voz incorrecto. Lo anterior se logra con la implementación de aprendizaje profundo generativo,

implementando adaptación de dominio y técnicas de filtrado basado en la transformada Wavelet. La integración de estas técnicas pretende mejorar la precisión y la calidad de la señal de voz, permitiendo una mayor confiabilidad en el reconocimiento del habla en entornos ruidosos.

1.4 Preguntas de investigación

1. ¿Qué tipos de modelos implementados existen en el estado de arte para la disminución de ruido en el reconocimiento de voz?
2. ¿En qué etapa de procesamiento de la señal de voz se debe aplicar la técnica de filtrado para disminuir el ruido contenido en la señal de voz?
3. ¿Por qué aplicar una etapa de preprocesamiento para la disminución de ruido sería adecuada, antes de someter a las características del habla al modelo acústico de aprendizaje profundo?
4. ¿En qué módulo de la arquitectura del sistema de RAV se acoplará la adaptación de dominio de la señal de voz para obtener un mejor reconocimiento del habla?
5. Dado que la adaptación de dominio consiste en transformar una distribución “fuente” a una distribución “objetivo”, ¿por qué el reconocimiento de voz tendrá un mejor rendimiento en el modelo de aprendizaje en diferentes entornos ruidosos?
6. ¿Qué resultados se obtendrán en la tarea de reconocimiento de voz en entornos ruidosos aplicando técnicas de filtrado y adaptación de dominio?, ¿y cuáles son sus ventajas y desventajas respecto a los modelos descritos en el estado del arte?

1.5 Objetivo general

Aplicar técnicas de filtrado y adaptación de dominio en la señal de voz con fines de reconocimiento del habla en entornos ruidosos para mejorar la tasa de reconocimiento.

1.6 Objetivos específicos

- 1 Analizar el efecto del ruido con varios tipos y niveles de ruido sobre el reconocimiento del habla, con base al estado del arte.
- 2 Definir el acoplamiento de técnicas de filtrado y adaptación de dominio en la arquitectura del reconocimiento del habla.
- 3 Evaluar las técnicas aplicadas en el reconocimiento del habla y analizar respecto a los trabajos relacionados de modelos de habla robustos contra el ruido.
- 4 Demostrar que los resultados obtenidos experimentales, aplicando técnicas de filtrado y adaptación de dominio mejoran al menos en un 5%, el robustecimiento del reconocimiento de voz en entornos ruidosos.

1.7 Hipótesis

H0: La aplicación de técnicas de filtrado y adaptación de dominio en la señal de voz mejoran al menos 5% el reconocimiento del habla respecto a los diversos entornos ruidosos.

1.8 Alcance

El presente trabajo propone desarrollar un modelo de aprendizaje profundo de reconocimiento de voz robusto, utilizando un preprocesamiento para disminuir el ruido en las señales de voz, aplicando técnicas de filtrado y una adaptación de dominio con transporte óptimo enfocado en el reconocimiento del habla.

1.9 Estructura de la tesis

El presente trabajo se organiza como se describe a continuación. En el Capítulo 2 nos enfocamos en descripción de las teorías base, las arquitecturas, y las contribuciones de los sistemas de reconocimiento automático de voz. En el Capítulo 3 se describe el modelo de investigación a seguir, la descripción de la propuesta, así como la selección de la muestra de las señales voz, y también cómo se realiza su posterior análisis bajo el enfoque propuesto. En el Capítulo 4 se analizan los resultados y limitaciones de la propuesta planteada. Y finalmente el Capítulo 5 se centra en los objetivos alcanzados, así como una discusión de la hipótesis puesta a prueba y contribuciones de la investigación.

Capítulo 2 Reconocimiento del habla

En este apartado se abordan las teorías fundamentales sobre los sistemas de reconocimiento del habla, centrándose específicamente en un reconocedor automático de voz y en la identificación de locutor.

2.1 Descripción de los sistemas del reconocimiento del habla

Un reconocedor del habla es un dispositivo que transcribe automáticamente el habla en texto o en un comando. El reconocedor generalmente se basa en un vocabulario finito que restringe las palabras que se pueden reconocer (Charniak, 1994). La arquitectura típica de un sistema reconocedor de voz se ilustra en la figura 2.1. Estos sistemas tienen cuatro componentes principales, que son i) procesamiento de señales, ii) extracción de características, modelo acústico, iii) modelo de lenguaje y iv) búsqueda de hipótesis. El componente de procesamiento de señales y extracción de características toma como entrada la señal de audio, mejora el habla mediante la eliminación de ruidos y distorsiones de canal, convierte la señal del dominio del tiempo al dominio de la frecuencia y extrae vectores de características. El modelo acústico integra conocimientos sobre acústica y fonética, toma como entrada las características generadas a partir del componente de extracción de características y genera una probabilidad del modelo acústico para la secuencia de características de longitud variable. El modelo de lenguaje estima la probabilidad de una secuencia de palabras hipotética, aprendiendo la correlación entre palabras de un corpus de entrenamiento (Yu & Deng, 2014).

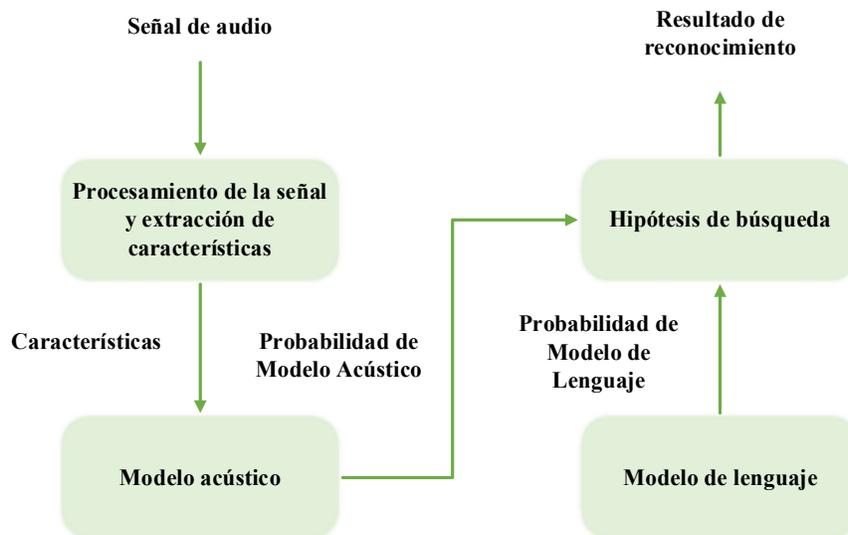


Figura 2.1 Arquitectura de un sistema reconocedor de voz

2.2 Principales desafíos de los reconocedores del habla

De acuerdo con los autores Yu & Deng (2014), un sistema de reconocimiento de voz exitoso debe lidiar con toda esta variabilidad acústica. A medida que pasamos de las tareas restringidas a

las aplicaciones del mundo real, como se ilustra en la figura 2.2, los sistemas de reconocimiento del habla, hoy en día, necesitan lidiar con un vocabulario enorme, conversaciones de estilo libre, habla espontánea ruidosa de campo lejano y mezcla de lenguajes.

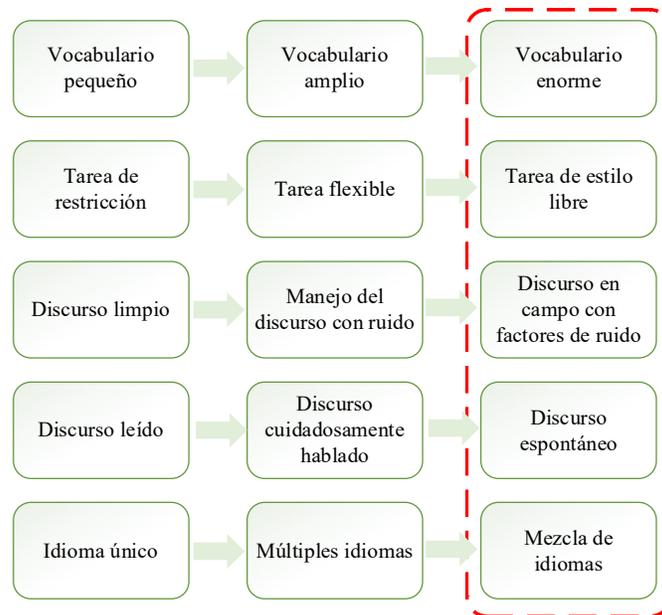


Figura 2.2 Los problemas actuales de un RAV (columna derecha) son mucho más difíciles que en los que hemos trabajado en el pasado debido a la demanda de las aplicaciones del mundo real

El reconocimiento del habla, de acuerdo con Yu & Deng (2014), interviene con cinco pasos principales en la secuencia básica:

- Adquisición de señales de voz: uso de micrófonos con cancelación de ruido o de conjuntos de micrófonos, técnicas de cancelación de ruido adaptativa o activa.
- Análisis y parametrización acústica.
- Segmentación y detección de voz y no voz o silencios: deben diseñarse técnicas sólidas, ya que una serie de errores de reconocimiento de voz con ruido se originan a partir de una determinación incorrecta de los límites de las expresiones.
- Modelado de patrones de referencia.
- Algoritmos de reconocimiento y medidas de distancia. Por supuesto, los diferentes métodos no son excluyentes y pueden combinarse para obtener rendimientos satisfactorios.

2.3 Métodos para el reconocimiento del habla

Existen diversas formas o enfoques para reconocimiento del habla, entre las cuales se incluyen métodos como el reconocimiento de patrones, sistemas basados en el reconocimiento mediante modelos estocásticos y el uso de redes neuronales.

2.3.1 Reconocimiento de patrones

La aproximación basada en el contraste de patrones supone que la oración formulada puede representarse como una secuencia de unidades del habla (palabras), cada una representada por un cierto patrón o conjunto de patrones. Se establece un criterio de distancia o similitud que permita comparar una unidad de habla con cada uno de los patrones de referencia almacenados y que sirva para determinar el patrón que mejor se ajuste, en algún sentido, a la unidad de entrada (Becerra, 2017; Haridas et al., 2018; Multisensor, 2014).

2.3.2 Sistemas basados en el conocimiento

Los sistemas basados en el conocimiento tratan de emular un conjunto de conocimientos sobre el habla, puesto en juego por un ser humano en su tarea de comprensión de un discurso. Para ello hace uso de técnicas de construcción de sistemas basados en reglas y sistemas expertos, desde el mismo nivel acústico-fonético, o bien desde niveles superiores (Becerra, 2017; Haridas et al., 2018; Multisensor, 2014).

2.3.3 Modelos estocásticos

Supone un avance en la capacidad de generalización. Una característica diferenciadora es la utilización de modelos probabilísticos en lugar de determinísticos, teniendo así capacidad de integración para una solución simultánea del problema de segmentación y el de clasificación (Becerra, 2017; Haridas et al., 2018; Multisensor, 2014).

2.3.4 Modelos neuronales o conexionistas

Alternativa capaz de realizar un trabajo similar a los modelos estocásticos sin tener tantas restricciones y tiempos aceptables (Becerra, 2017; Haridas et al., 2018; Multisensor, 2014).

2.4 Descripción de las teorías base de un sistema de reconocimiento de voz

Para el diseño de un reconocedor de voz se necesita de una formulación matemática. Para tratar este problema es necesario descomponer el problema en subproblemas para un mejor manejo dentro del reconocedor de voz. Se han propuesto (Charniak, 1994) formulaciones matemáticas para el manejo del sistema de reconocimiento automático de voz, que a continuación se describen.

2.4.1 Formulación matemática

Sea A la evidencia acústica sobre la base de la cual el reconocedor tomará su decisión sobre las palabras que se pronunciaron, A es una secuencia de símbolos tomados de algún alfabeto \mathcal{A} en particular y está dada por la ecuación (2.1).

$$A = a_1, a_2, \dots, a_m \quad a_i \in \mathcal{A}. \quad (2.1)$$

Los símbolos a_i se tratan como lo que se ha generado en el tiempo, como lo indica el índice i . Sea también la ecuación (2.2) una cadena de n palabras, de la cual cada palabra pertenece a un vocabulario fijo conocido \mathcal{V} .

$$W = w_1, w_2, \dots, w_m \quad w_i \in \mathcal{V}. \quad (2.2)$$

Si $P(W|A)$ denota la probabilidad de que se pronuncien palabras W , dado que se observó la evidencia A , entonces el reconocedor debe decidir a favor de una cadena de palabras W que satisfaga la ecuación (2.3).

$$\hat{W} = \underset{w}{\arg \max} P(W|A). \quad (2.3)$$

Es decir, el reconocedor elegirá la cadena de palabras más probable dada la evidencia acústica observada. La fórmula basada en la teoría de la probabilidad de Bayes nos permite reescribir la probabilidad de la ecuación (2.3) como se muestra la ecuación (2.4).

$$P(W|A) = \frac{P(W)P(A|W)}{P(A)}, \quad (2.4)$$

donde $P(W)$ es la probabilidad a priori de que se pronuncie la cadena de palabras W , $P(A|W)$ es la probabilidad de que cuando el hablante dice W se observe la evidencia acústica A , y $P(A)$ es la probabilidad media de que A será observado. Esto está dado por la ecuación

$$P(A) = \sum_{w'} P(W') P(A|W'). \quad (2.5)$$

Dado que la maximización en (2.3) se lleva a cabo con la variable A fija, se deduce de (2.3) y (2.4) que el objetivo del reconocedor es encontrar la cadena de palabras \hat{W} que maximiza el producto $P(W)P(A|W)$, es decir, satisface la ecuación (2.6).

$$\hat{W} = \underset{w}{\arg \max} P(W)P(A|W). \quad (2.6)$$

La fórmula (2.6) determina qué procesos y componentes son importantes en el diseño de un reconocedor de voz.

2.4.2 Procesamiento acústico

Primero, es necesario decidir qué datos acústicos A se observarán. Es decir, se necesita decidir por un extremo frontal que transformará la forma de onda de presión (que es el sonido) en los símbolos a_i con los que tratará el reconocedor. Entonces, en principio, esta interfaz incluye un micrófono cuya salida es una señal eléctrica, un medio para muestrear esa señal y una forma de procesar la secuencia de muestras resultante (Charniak, 1994).

2.4.3 Modelado acústico

El reconocedor debe ser capaz de determinar el valor $P(A|W)$ de la probabilidad de la ecuación (2.6) de que cuando el hablante pronuncie la secuencia de palabras W , y el procesador acústico

produzca los datos A . Dado que este número debe estar disponible para todos los posibles emparejamientos de W con A , se deduce que debe ser computable o calculado "sobre la marcha". El número de diferentes valores posibles de A y W es demasiado grande para permitir una búsqueda. Por lo tanto, para calcular $P(A|W)$ necesitamos un modelo acústico estadístico de la interacción del hablante con el procesador acústico. El proceso total que estamos modelando implica la forma en que el orador pronuncia las palabras de W , el ambiente (ruido de la sala, reverberación, etc.), la ubicación y las características del micrófono, y el procesamiento acústico realizado por el front-end (Charniak, 1994).

2.4.4 Modelo de lenguaje

La fórmula (2.6) requiere calcular para cada cadena de palabras W la probabilidad a priori $P(W)$ de que el hablante desee pronunciar W . La fórmula de Bayes permite muchas descomposiciones de $P(W)$, pero debido a que el reconocedor desea transmitir el texto en la secuencia en que fue pronunciado, se utiliza la descomposición de la ecuación (2.7):

$$P(W) = \prod_{i=1}^n P(w_i | w_1, \dots, w_{i-1}). \quad (2.7)$$

Por tanto, el reconocedor debe poder determinar estimaciones de las probabilidades $P(w_i | w_1, \dots, w_{i-1})$. Se usa el término estimación a propósito, porque incluso para valores moderados de i y vocabularios de tamaño razonable, la probabilidad $P(w_i | w_1, \dots, w_{i-1})$ tiene demasiados argumentos. La fórmula (2.7) en realidad puede realizarse como la ecuación (2.8):

$$P(W) = \prod_{i=1}^n P(w_i | \phi(w_1, \dots, w_{i-1})). \quad (2.8)$$

y el modelado del lenguaje consiste en determinar la clasificación de equivalencia adecuada y un método para estimar las probabilidades $P(w_i | w_1, \dots, w_{i-1})$. El modelo de lenguaje utilizado debe depender del uso que se le dará al reconocedor. El modelo de lenguaje dependerá solo del texto y de ninguna manera del habla (Charniak, 1994).

2.4.5 Hipótesis de búsqueda

Para encontrar la transcripción deseada \hat{W} de los datos acústicos A por la fórmula (2.6), se debe buscar entre todas las cadenas de palabras posibles W para encontrar la máxima probabilidad. Esta búsqueda no puede llevarse a cabo por la fuerza bruta: el espacio de W_s es astronómicamente grande. Por lo tanto, se necesita una búsqueda de hipótesis que ni siquiera considerará el abrumador número de posibles candidatos W y examinará solo aquellas cadenas de palabras sugeridas de alguna manera por la acústica A (Charniak, 1994).

2.5 Esquemas de reconocimiento de voz

Existen dos enfoques clásicos ampliamente usados en arquitecturas de sistemas de reconocimiento automático de voz, 1) los modelos híbridos basados en Modelos de Mezclas

Gaussianas y 2) los basados en Redes Neuronales Artificiales Profundas y más recientemente se ha propuesto el enfoque End-to-End. Los componentes principales en las dos arquitecturas de un reconocedor de voz son el componente de extracción de características en el front-end y la clasificación en el back-end. En la fase de extracción de características, la información que es importante para el reconocimiento correcto se recoge en una secuencia discreta de vectores de características. Estas características se procesan para encontrar unidades lingüísticas como palabras, sílabas y fonos. En el back-end, los modelos acústicos predicen la puntuación de coincidencia para cada palabra.

2.5.1 Modelos híbridos - Mezclas Gaussianas – Modelos Ocultos de Markov

El modelo híbrido MMG-MOM, de acuerdo con Bansal et al. (2008), tiene la capacidad de encontrar la máxima probabilidad conjunta entre todas las posibles palabras de referencia W dada la secuencia de observación O . En el caso real, la combinación de los MMG y los MOM con un coeficiente ponderado puede ser un buen esquema debido a la diferencia en los métodos de entrenamiento. El i -ésimo MMG independiente del hablante produce probabilidad $P^i(MMG)$, $P = 1, 2, \dots, W$, donde W es el número de palabras. El i -ésimo MOM independiente del hablante también produce probabilidad $P^i(MOM)$, $P = 1, 2, \dots, W$. Todos estos valores de probabilidad se pasan al llamado bloque de decisión de probabilidad, donde se transforman en la nueva probabilidad combinada $P'(W)$:

$$P'(W) = (1 - x(W))P^i(GMM) + x(W)P^i(MOM), \quad (2.9)$$

donde $x(W)$ denota un coeficiente de ponderación. El valor de x se calcula durante el entrenamiento del modelo híbrido. En las pruebas híbridas, se utiliza el subconjunto de datos de entrenamiento y se calculan los valores de probabilidad de MOM y MMG (ver el esquema en la figura 2.3), que se combinan mediante el coeficiente de ponderación. Los valores estáticos del coeficiente ponderado también se utilizan para obtener una mayor tasa de reconocimiento.

2.5.2 Redes Neuronales Artificiales – Modelos Ocultos de Markov

De acuerdo con Willett & Rigou, (1997), la arquitectura básica de este sistema híbrido se ilustra en la figura 2.4. La red neuronal funciona como una transformación de características que tiene en cuenta varios vectores de características pasados y futuros adicionales para producir un vector de características mejorado y más discriminatorio que se alimenta al sistema con MOM. Esta arquitectura permite tres formas de interpretación:

1. Como un sistema híbrido que combina redes neuronales y MOM continuos.
2. Como una transformación similar que incorpora los parámetros MOM en el cálculo de la matriz de transformación.
3. Como el método de extracción de características que permite este proceso de acuerdo con el sistema MOM subyacente.

Los tipos de redes neuronales considerados son transformaciones lineales, red neuronal artificial y red neuronal artificial recurrente. Con esta arquitectura, se pueden tener en cuenta vectores de características pasadas y futuras adicionales en el proceso de estimación de probabilidad sin aumentar la dimensionalidad de los componentes de la mezcla gaussiana. En lugar

de aumentar el número de parámetros del sistema MOM, la red neuronal se entrena para producir vectores de características más discriminantes con respecto al sistema MOM entrenado. Por supuesto, agregar algún tipo de red neuronal también aumenta la cantidad de parámetros, pero el aumento es mucho más moderado de lo que sería al aumentar la dimensionalidad de cada gaussiana.

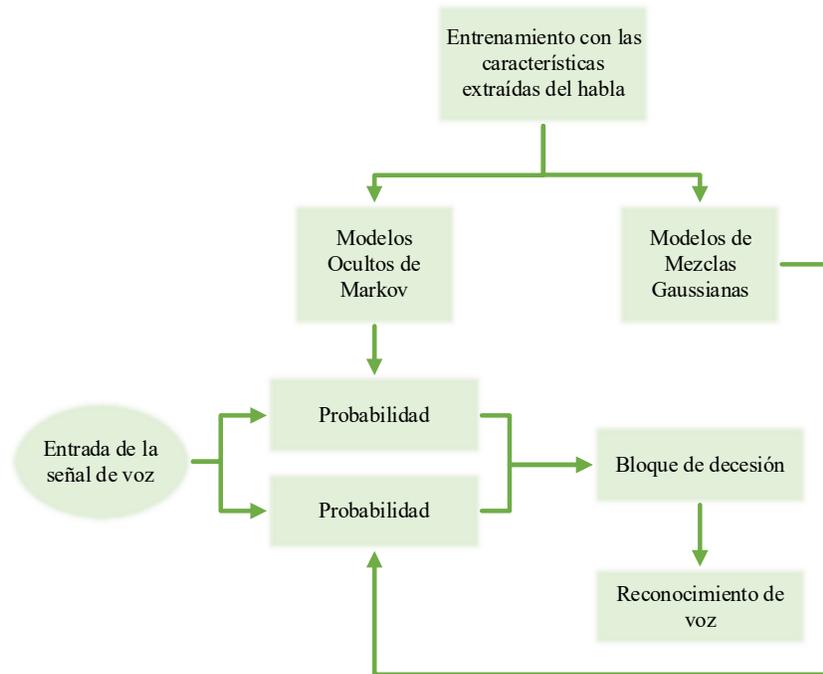


Figura 2.3 Arquitectura de un sistema de reconocimiento automático de voz MMG-MOM

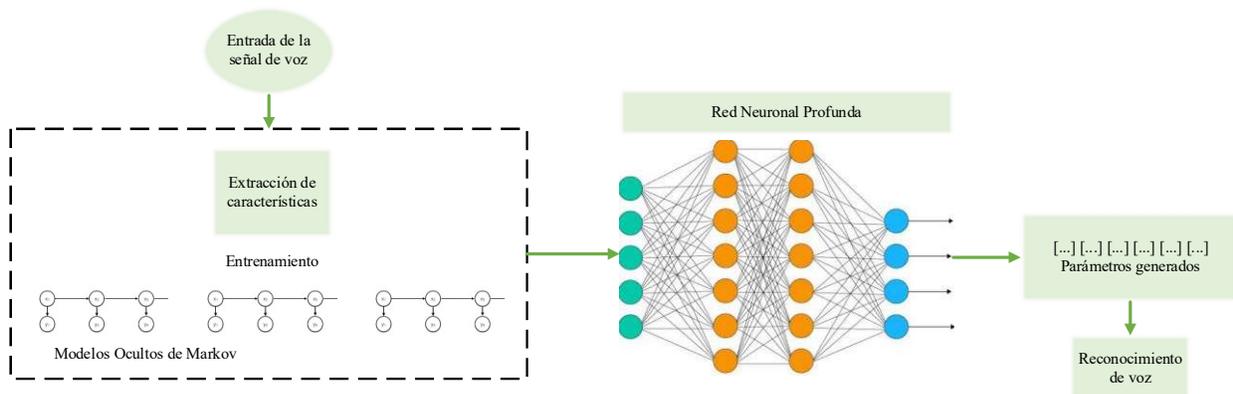


Figura 2.4 Arquitectura de un sistema de reconocimiento automático de voz RNP-MOM

2.5.3 Modelos End-to-End

De acuerdo con (S. Wang & Li, 2019) un modelo End-to-End es un sistema que mapea

directamente una secuencia de características acústicas de entrada en una secuencia de grafemas o palabras. Un sistema está entrenado para optimizar los criterios que están relacionados con la métrica de evaluación final que es la tasa de error de palabras. Para un reconocedor automático de voz convencional, la mayoría de los sistemas de reconocedor automático de voz involucran componentes de modelos acústicos, de pronunciación y de lenguaje entrenados por separado; selección del léxico de pronunciación, la definición de fonemas para el idioma en particular, además de requerir conocimiento experto y de mucho tiempo. La figura 2.5 representa la estructura de un reconocedor automático de voz End-to-End.

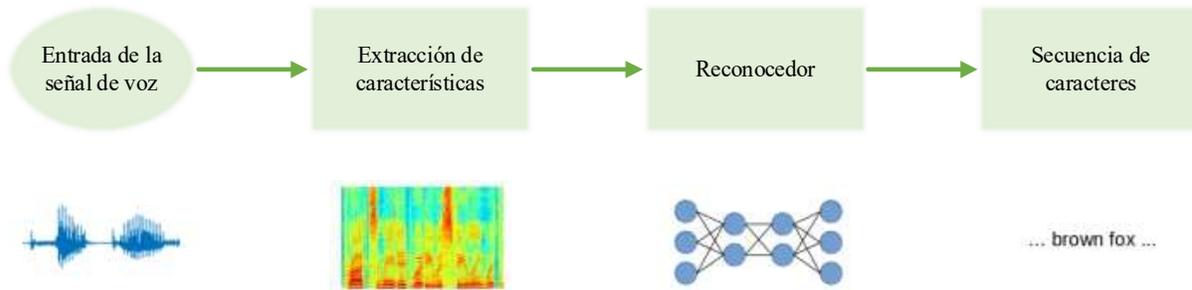


Figura 2.5 Arquitectura de un sistema de reconocimiento automático de voz End-to-End

Se puede ver que el reconocimiento de voz del esquema End-to-End simplifica enormemente la complejidad respecto a la del reconocimiento de voz híbrido. No hay necesidad de etiquetar manualmente la información, la red neuronal puede aprender automáticamente el idioma o la información de pronunciación, como se muestra en la figura 2.5. Ahora hay dos estructuras principales para el reconocimiento de voz de End-to-End, que son la Clasificación Temporal Conexionista (CTC) y el modelo de atención.

2.6 Descripción de las teorías base de un sistema para identificación de locutor

En esta sección se presentan las teorías fundamentales para llevar a cabo la identificación del locutor, abarcando desde la extracción de características hasta el modelado para el aprendizaje de la identificación del locutor.

2.6.1 Identificación de locutor

La identificación de locutor es el proceso de distinguir consecuentemente al individuo que habla en la muestra de voz. Es la tarea más desafiante debido a que cada hablante es diferente en términos de acento, estilo de habla, frecuencia de palabras y tracto vocal. La presencia de ruido, conversaciones de fondo y música también dificulta aún más la tarea (Ashar et al., 2020). Las condiciones como un dispositivo de grabación defectuoso también afectan la precisión de la clasificación. En la identificación del hablante en un entorno cerrado y en configuración de independencia de texto, la voz debe ser de un hablante inscrito y no depende de las palabras dichas por el hablante. La voz es probablemente el modo de contacto más importante entre los seres humanos. De acuerdo con la literatura expuesta por los autores Anggun et al. (2018); Hidayat &

Winursito (2020); Q. Li et al. (2018, 2020); Risanuri et al. (2018) en sus trabajos desarrollados, se menciona que la extracción de vectores de coeficientes cepstrales de frecuencia de Mel han sido ampliamente utilizados como características para los sistemas de reconocimiento de voz automático. Ellos mencionan que los MFCC se han convertido en un método popular para la descripción de la señal de voz. Cabe señalar que para la extracción de características de una señal de voz se suelen utilizar otros métodos para el procesamiento del sistema de reconocimiento de patrones, en el que involucran la señal de voz. Además, la extracción de MFCC está inspirada en el mecanismo auditivo humano, donde añadiendo a lo anteriormente comentado, los MFCC se convierten en la función más utilizada; además que ha demostrado tener una precisión durante el modelado de estos (Jurafsky & Martin, 2000; Q. Li et al., 2020). Por lo que en las secciones siguientes se describirán los pasos involucrados para la extracción de vectores MFCC.

2.6.2 Procesamiento acústico

La voz humana o el habla es una señal rica en información que transmite una amplia gama de datos, como el contenido del lenguaje, las emociones del hablante y el tono del discurso. El reconocimiento de patrones de voz busca separar, identificar y reconocer a un hablante basándose en características del habla. Varios métodos pueden simplificar el proceso de reconocimiento del hablante. Estos sistemas generalmente implican dos fases: i) extracción de características y ii) coincidencia o clasificación de las características, donde el componente de clasificación tiene dos elementos: a) la coincidencia de patrones y b) la toma de decisiones (Bunrit et al., 2019).

2.6.3 Extracción de características

El módulo de extracción de características estima una colección de características de la señal de habla que reflejan información específica del hablante, donde la voz de cada hablante se recopila y se utiliza para construir el modelo correspondiente del hablante. La compilación de modelos de voz para todos los hablantes se denomina conjunto de datos de voz. El proceso de extracción de características basado en coeficientes MFCC se muestra en la figura 2.6. El método comienza con la normalización y el preprocesamiento de la señal de habla o acústica si es necesario. Luego, se realiza el análisis temporal mediante el uso de operaciones de enmarcado y ventaneo antes del proceso de la transformada de Fourier. Después de eso, se aplica un banco de filtros en escala Mel y se envuelve en una escala logarítmica. A continuación, se utiliza la Transformada Coseno Discreta como un proceso antes de calcular la Sustracción Media Cepstral. Estos coeficientes son las características extraídas mediante el método de coeficientes MFCC (Khdier et al., 2021). Esto significa que el método de extracción de características basado en MFCC utiliza el análisis de frecuencia basado en el procesamiento de la señal de habla a través del banco de filtros. El resultado del método son las características extraídas en forma de coeficientes MFCC. Estos coeficientes MFCC, como resultado, pueden ser utilizados posteriormente para el análisis o clasificación con cualquier propósito (Bunrit et al., 2019).

El objetivo principal de esta sección es describir los bloques de la figura 2.6, esto es, cómo podemos transformar la señal de voz del locutor para obtener sus características representativas y una secuencia de características de vectores acústicos (Jurafsky & Martin, 2000).

2.6.3.1 Señal de voz o señal acústica

Primero es necesario decidir las observaciones de datos para el procesamiento acústico; eso es necesario y es lo primero que se lleva a cabo para transformar la forma de onda de voz del locutor (esta es la que contiene el sonido), y esto es lo que el reconocedor de voz procesa. En principio durante esta etapa mediante un micrófono se convierte en una señal eléctrica, lo que significa muestrear la señal y almacenarla en un medio de almacenamiento (Ashar et al., 2020).

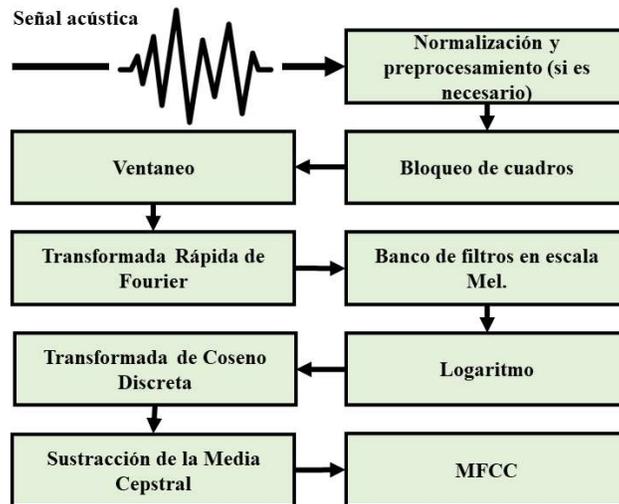


Figura 2.6 Extracción de MFCCs

2.6.3.2 Obtención de la señal de voz

Este proceso consiste en muestrear y cuantizar la forma de onda de voz analógica. Durante la etapa de muestreo, la señal de voz es muestreada midiendo su amplitud en un tiempo en particular. La tasa de muestreo es el número de muestras tomadas por segundo. El número de muestras tomada por segunda se rige por la frecuencia de Nyquist. La información de la señal de voz humana está debajo de las frecuencias 10,000 Hz, por lo tanto, una tasa de muestreo de 20,000 es necesaria para cumplir con las condiciones del teorema de muestreo uniforme. Pero hablando de la voz transmitida por vía telefónica, ésta es filtrada para red de conmutación, y sólo las frecuencias menores de 4000 Hz son transmitidas a través de telefonía. Por lo tanto, una frecuencia de muestreo de 8000 Hz es suficiente para la transmisión de la voz. El almacenamiento de la señal digital usualmente se almacena en espacios de memoria (enteros), ya sea de 8 bits (valores de -128 - 127) o 16 (valores de -32768 - 32767). Este proceso de representar valores reales como enteros es llamado cuantización, dado que hay una granularidad mínima y todos los valores que están más cerca entre sí que este tamaño cuántico se representan de forma idéntica (Jurafsky & Martin, 2000).

2.6.3.3 Pre-énfasis

El primer paso en la extracción de características MFCC es aumentar la cantidad de energía en las frecuencias altas (Jurafsky & Martin, 2000).

2.6.3.4 Ventaneo

El habla es una señal no estacionaria, esto quiere decir que sus propiedades estadísticas no son constantes a lo largo del tiempo. La extracción de la señal se realiza multiplicando el valor de la señal en el instante n , $s[n]$, por el valor de la ventana en el instante n , $w[n]$ tal como se describe en la ecuación (2.10) (Jurafsky & Martin, 2000).

$$y[n] = w[n]s[n]. \quad (2.10)$$

Para ello se realiza un ventaneo y así convertir una señal no estacionaria a una señal cuasi estacionaria para extraer sus características espectrales de una pequeña ventana de voz que caracteriza un sub-fono particular. Este proceso de creación de ventanas se caracteriza mediante 3 parámetros: I) el ancho de la ventana (en milisegundos), II) el desplazamiento entre las ventanas sucesivas y III) finalmente el marco de la ventana; este puede ser rectangular, dado por la ecuación (2.11), o Hamming, dado por la ecuación (2.12) (Jurafsky & Martin, 2000).

$$\text{Rectangular} \quad w[n] = \begin{cases} 1, & 0 \leq n \leq L - 1 \\ 0, & \text{De otra manera} \end{cases} \quad (2.11)$$

$$\text{Hamming} \quad w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L}\right), & 0 \leq n \leq L - 1 \\ 0, & \text{De otra manera} \end{cases} \quad (2.12)$$

2.6.3.5 Transformada Discreta de Fourier

El siguiente paso es extraer, mediante la Transformada Discreta de Fourier, la información espectral por cada ventana generada del paso anterior. Está extrae la información necesaria sobre la energía contenida de la señal en las diferentes bandas de la frecuencia. La entrada para este proceso es una ventana de la señal de voz y la salida será para cada uno de los valores discretos de las bandas de frecuencia un número complejo que representa la magnitud y fase de ese componente de la frecuencia en la señal original. La Transformada Discreta de Fourier está definida por la ecuación (2.13) (Jurafsky & Martin, 2000).

$$X(k) = \sum_{n=0}^{N-1} x[n]e^{-j2\frac{\pi}{N}kn} \quad (2.13)$$

2.6.3.6 Banco de filtros Mel y Log

Los resultados obtenidos por la Transformada Discreta de Fourier brindarán información sobre la cantidad de energía para cada banda de frecuencia. Sin embargo, el oído humano no tiene la capacidad para percibir todas las bandas de frecuencia, este es menos sensible a frecuencias más altas. Es por ello, que modelar esta propiedad de la audición humana mejorará el rendimiento del reconocimiento de voz. Este proceso consiste en deformar las salidas de frecuencias de la Transformada Discreta de Fourier a una escala mel. Un mel es una unidad de tono definida de modo que los pares de sonidos que son perceptivamente equidistantes en tono están separados por un número igual de mels. El mapeo entre la frecuencia en Hertz y la escala de mel es lineal por

debajo de 1000 Hz y logarítmico por encima de 1000 Hz. La frecuencia mel m se puede calcular a partir de la frecuencia acústica bruta mediante la ecuación (2.14) (Jurafsky & Martin, 2000).

$$mel(f) = 1127 \ln \left(1 + \frac{f}{700} \right). \quad (2.14)$$

2.6.3.7 El Cepstrum: Transformada inversa discreta de Fourier

El Cepstrum tiene un número de ventajas útiles y mejora significativamente el rendimiento del reconocimiento de un fonema. El Cepstrum consiste en separar la fuente glótica y el filtro. La forma de onda del habla se crea cuando una forma de onda de fuente glótica de una frecuencia fundamental particular pasa a través del tracto vocal, que debido a su forma tiene una característica de filtrado particular. Pero las características de la fuente glótica no son importantes para distinguir los diferentes fonemas. Por otro lado, la información más útil para la detección de fonemas es el filtro, ya que este proporciona la posición exacta del tracto vocal, por lo que se detectaría el tipo de fonema. El Cepstrum comienza con un espectro de magnitud estándar, luego se reemplaza cada amplitud de valor en el espectro de magnitud con su logaritmo. Al realizar el proceso se obtendrá la frecuencia fundamental que representa el pulso glótico y los demás componentes que representan el filtro del tracto vocal. El Cepstrum se define formalmente en la ecuación (2.15) como la inversa de la magnitud logarítmica de la Transformada Inversa de Fourier de una señal, por lo tanto, para una ventana marco de expresión $x[n]$ (Jurafsky & Martin, 2000).

$$c[n] = \sum_{k=0}^{N-1} \log \left(\left| \sum_{n=0}^{N-1} x[n] e^{-j \frac{\pi}{N} kn} \right| \right) e^{-j 2 \frac{\pi}{N} kn}. \quad (2.15)$$

2.6.3.8 Energía y deltas

La energía se correlaciona con la identidad del fonema y por lo tanto es una característica para la identificación de fonemas. La energía en una ventana es la suma en el tiempo de la potencia de muestras contenidas en la ventana. Una señal dada en una ventana, desde la muestra de tiempo t_1 hasta la muestra t_2 , tiene una energía que matemáticamente está definida por la ecuación (2.16) (Jurafsky & Martin, 2000).

$$Energía = \sum_{n=0}^{N-1} x^2[t]. \quad (2.16)$$

Otro factor importante para la señal de voz a tomar en cuenta radica en que no es constante de una ventana a otra. Este cambio, tal como la pendiente de una formante en su transición, o el cambio natural de una ventana a otra, puede proporcionar características útiles para la identificación de fonemas. Para extraer la característica de ventana a ventana se añade por cada característica (energía y Coeficientes Cepstrales de Frecuencia Mel de la ventana) una delta o característica de velocidad, la cual corresponde al cambio entre ventanas de las características cepstrales y energía; además se añade una doble delta o característica de aceleración que representa el cambio y entre ventanas entre las características delta. Las deltas y dobles deltas pueden ser calculadas con la ecuación (2.17), donde $d(t)$ se obtiene para un valor cepstral particular, $c(t)$ en el tiempo t (Jurafsky & Martin, 2000).

$$d(t) = \frac{c(t+1) - c(t-1)}{2}. \quad (2.17)$$

2.6.4 Modelos para identificación de locutor

Existen varios algoritmos utilizados para la identificación de locutor, cada uno con sus propias características y enfoques (Rosenberg, 1976). Algunos de los algoritmos más comunes son:

- Los modelos de Vector de Características *i*-vectores han demostrado ser efectivos en la identificación del locutor, especialmente en escenarios de grandes bases de datos y donde se busca una representación global del locutor en lugar de características locales en el habla. Estos modelos pueden manejar variaciones del habla a largo plazo y han sido ampliamente utilizados en aplicaciones de reconocimiento y autenticación de locutores.
- Los modelos basados en técnicas clásicas de aprendizaje máquina se utilizan para clasificar y distinguir a los locutores basándose en las características extraídas de las muestras de voz.
- Los modelos de aprendizaje profundo son capaces de aprender automáticamente las representaciones de las características del habla y realizar la identificación del locutor. Pueden manejar tanto características acústicas a nivel de trama como secuencias temporales más largas.

Capítulo 3 Adaptación de la señal del habla para reducción de ruido

En este capítulo se aborda la adaptación de la señal del habla con el objetivo de reducir el ruido. En primer lugar, se proporcionan conceptos generales sobre el ruido con el fin de comprender el problema en cuestión. Además, se presentan las teorías específicas que se utilizarán en este trabajo para abordar la reducción del ruido en la señal de voz. Por último, se discuten los trabajos relacionados que respaldan la metodología de investigación propuesta en este estudio.

3.1 Ruido

El ruido en señales son las modificaciones no deseadas que una señal puede sufrir durante la captura, almacenamiento, transmisión, procesamiento o conversión de esta (Singh & Singh, 2015). Estos factores contribuyen a alterar las propiedades originales de la señal, generando alteraciones no deseada en su forma. Este efecto en la señal puede afectar la calidad y la inteligibilidad de está, lo que implica la necesidad de abordar estos problemas para mejorar la percepción de la señal de interés.

3.1.1 Tipos de ruido

Hay dos tipos de ruido que son: estacionario y no estacionario.

3.1.1.1 *Ruido estacionario*

El término estacionario se refiere a que las características estadísticas del ruido, como la intensidad, la forma del espectro u otros factores, no experimentan cambios a lo largo del tiempo. En otras palabras, el concepto de estacionario implica que ninguno de los parámetros estadísticos del proceso sufre desplazamientos en su posición dentro del espacio de parámetros. En el contexto del ruido, esto implica que sus propiedades se mantienen constantes, lo que permite un análisis más predecible y una mejor comprensión de su comportamiento (Singh & Singh, 2015).

3.1.1.2 *Ruido no estacionario*

El ruido no estacionario se refiere a un tipo de ruido cuyas características estadísticas cambian a lo largo del tiempo. A diferencia del ruido estacionario, en el que las propiedades del ruido son constantes, el ruido no estacionario presenta variaciones en la intensidad, la forma del espectro u otros parámetros estadísticos a medida que transcurre el tiempo. Esto significa que las propiedades del ruido no se mantienen constantes y pueden cambiar, lo que puede dificultar su análisis y la implementación de estrategias para reducir su impacto en una señal de interés. El ruido no estacionario puede ser causado por diferentes factores, como cambios en las condiciones ambientales, interferencias eléctricas o variaciones en la fuente de la señal (Singh & Singh, 2015).

3.2 Adaptación de dominio

La adaptación de dominio es el proceso de entrenar una red neuronal en un conjunto de datos

fuente y lograr una buena precisión en el conjunto de datos objetivo, el cual es significativamente diferente al conjunto de datos fuente (Wu & He, 2022). Los métodos de adaptación de dominio entrenan un modelo para encontrar representaciones de características similares entre un dominio fuente y un dominio objetivo (Baffour et al., 2022). Los métodos recientes aprovechan el aprendizaje para descubrir representaciones análogas de los dos dominios. Existen varios enfoques tales como menciona el autor Fatras et al., (2021):

- Configuración de adaptación de dominio estándar. Ambos dominios comparten el mismo espacio de etiquetas $Y_s = Y_t$.
- Configuración de adaptación de dominio no supervisado. Se tiene un conjunto de datos de dominio de origen etiquetado $D_s = \{(x_i^s, y_i^s)\}_{i=1}^n, x_i \in \mathbb{R}^d$ y un conjunto de datos de dominio de destino sin etiquetar $D_t = \{(x_j^t)\}_{j=1}^n, x_j \in \mathbb{R}^d$.
- Configuración de adaptación de dominio semi-supervisado. Se tienen disponibles algunas etiquetas en el dominio de destino y la adaptación de dominio supervisado donde todas las etiquetas en el dominio de destino están disponibles.
- Configuración de adaptación de dominio de múltiples fuentes. Se tiene acceso respectivamente a varios conjuntos de datos de origen o de destino.

3.2.1 Introducción a la teoría de Transporte Óptimo para la adaptación de dominio

El Transporte Óptimo (Chehebar & Groisman, 2021; Harchaoui, 2020) está definido de manera informal (de acuerdo con Monge, quien lo propuso en Francia durante el siglo XVIII), como: “Considere que un obrero tiene una pala y debe mover una montaña de arena con una forma dada. Su objetivo será mover esa montaña de arena para construir en otro lugar una nueva forma deseada”. Para construir esa nueva forma existen múltiples maneras de llevar a cabo la actividad, pero se desea minimizar el esfuerzo, el cual se cuantificará como la distancia que se ha recorrido cargando la arena. Se puede normalizar la masa a uno, de forma que se representen distribuciones de probabilidad a las que se llamarán distribuciones (ver figura 3.1). Monge pensó en la distribución fuente (source: en rojo) como masa que habría que mover a otro lugar de forma que quede como lo indica la distribución objetivo (target: en azul). Lo que se quiere minimizar es el costo de transporte, que está dado por una función $c(x, y)$, que indica el costo de mover una masa de x a y .

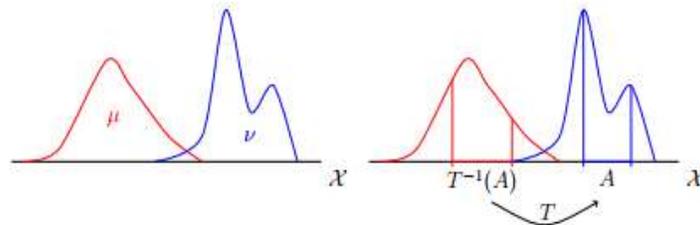


Figura 3.1 La imagen de la izquierda muestra dos medidas μ y ν en X dadas como densidades. La derecha muestra un subconjunto medible $A \subseteq X$ y su imagen inversa bajo un mapa de transporte T

Siguiendo ahora con las aplicaciones, se tiene que el campo de investigación del Transporte Óptimo (TO) ha demostrado ser crucial para redefinir nuestro mundo moderno tal como lo

conocemos, esto a través de una sorprendentemente amplia gama de aplicaciones, desde su nacimiento en Francia en el siglo XVIII, con científicos de muy diferentes formaciones y aplicaciones. Áreas como en procesamiento de imágenes, en una adaptación de dominio de datos casi en general, para la gestión de aeropuertos, para analizar victorias decisivas en batallas militares de la Segunda Guerra Mundial, innovación revolucionaria en la economía moderna y asombrosas revoluciones industriales futuristas como semi-diseño generativo de objetos automáticos (Harchaoui, 2020; Peyré, 2018; Santambrogio, 2018).

3.2.2 Tipos de Transporte Óptimo

Dependiendo de las medidas μ y ν , el problema de Transporte Óptimo puede tomar diferentes formas. Por lo general, se diferencia entre tres tipos de problemas (Schrieber, 2019):

- Transporte continuo. Ambas medidas son continuas.
- Transporte semi-discreto. Una de las medidas es continua, y la otra es discreta.
- Transporte discreto. Ambas medidas son discretas.

Estos tres tipos de problemas requieren métodos diferentes y cuidadosamente adaptados, y los algoritmos adecuados para resolver uno de estos tipos generalmente no se transfieren fácilmente a otros tipos. Sin embargo, es posible reformular un problema como de otro tipo.

3.2.3 Formulaciones del Transporte Óptimo

Dado un espacio de datos X con una métrica c , se quiere medir una distancia entre dos distribuciones de datos relacionados con esa métrica. Matemáticamente, se manipulan dos distribuciones μ y ν con dos variables asociadas $x \sim \mu$ y $y \sim \nu$, ambas en X . En esas distribuciones, queremos calcular la cantidad $T(x)$ midiendo qué tan diferentes son las pilas μ y ν . Para ello, existen tres formulaciones equivalentes para esa misma cantidad (Harchaoui, 2020).

3.2.3.1 Formulación de Monge

La formulación de Monge está dada por la ecuación (3.1), y lo que se quiere minimizar es el costo de ese transporte, que está dado por una función $c(x, y)$, que indica el costo de mover masa de x a y . El problema podría formularse entonces como minimizar el costo de transportar una distribución μ a una ν (ambas conocidas) sujeto a que a cada punto x se le asigna un $T(x)$ que es su transportado (Chehebar & Groisman, 2021).

$$\min_{T_{\#}\mu = \nu} \int c(x, T(x)) d\mu, \quad (3.1)$$

donde $T_{\#}\mu = \nu$ se refiere a que después de transportar con T la distribución μ obtendremos ν . Matemáticamente esto significa que ν es la medida push-forward de μ vía T . Típicamente esta función de transporte se define entre dos puntos de \mathbb{R}^d o bien entre dos espacios X y Y , que son los espacios en los que definimos las medidas μ y ν respectivamente, que en general serían subconjuntos de \mathbb{R}^d (no necesariamente ambos con el mismo d). La función T ofrece un transporte punto a punto, es decir, que conceptualmente toda la masa que hay en x se moverá a un único punto $y = T(x)$.

3.2.3.2 Formulación de Kantorovich

Por otro lado, se tiene la formulación de Kantorovich, se puede pensar que la masa del punto x está yendo a y , donde se imponen las restricciones de la ecuación (3.2) que nos dice que todo lo que sale de x es $\mu(x)$ y que todo lo que llega a y es $\nu(y)$ (Chehebar & Groisman, 2021).

$$\mu(x) = \int \pi(x, y) dy \qquad \nu(y) = \int \pi(x, y) dx. \qquad (3.2)$$

Se denominará $\Pi_{\mu, \nu}$ al conjunto de todas las distribuciones conjuntas que cumplen esto (Ambrosio & Gigli, 2009; Chehebar & Groisman, 2021). La formulación de Kantorovich del problema de Transporte Óptimo es entonces la siguiente:

$$\min_{\pi \in \Pi_{\mu, \nu}} \int c(x, y) d\pi(x, y). \qquad (3.3)$$

La formulación de Kantorovich se trata de una relajación de la de Monge porque permite que la masa de x se transporte a distintos lugares y no sólo a uno específico $T(x)$. Más precisamente, toda función T que cumpla las restricciones de la formulación de Monge induce una distribución conjunta $\pi(x, y) = \mu(x) \delta T(x)(y)$ que pertenece a $\Pi_{\mu, \nu}$ ya que cumple (3.2). En general, diremos que una solución es factible si cumple las restricciones pedidas, es decir $\pi \in \Pi_{\mu, \nu}$ en la formulación de Kantorovich y $T_{\#}\mu = \nu$ en la formulación de Monge.

3.2.3.3 Formulación de Wasserstein

Un aspecto muy importante del Transporte Óptimo es que da lugar a la distancia de Wasserstein, que está dada por la ecuación (3.4); está permite cuantificar el costo de trasladar una distribución en otra. A partir de una función de costo punto a punto, podemos obtener un costo de transportar una distribución en otra. En el caso discreto, podemos vía TO generalizar un costo punto a punto a un costo entre histogramas o conjuntos de puntos, que representan distribuciones discretas. Tener un costo entre distribuciones es muy útil ya que nos permite cuantificar cuáles de ellas se parecen y cuáles no (Chehebar & Groisman, 2021). La distancia de Wasserstein tiene varios nombres diferentes, como distancia del motor de la tierra, distancia de Mallows, distancia de Monge-Kantorovich-Rubinstein, dependiendo del campo en el que se utilice. (Schrieber, 2019).

$$W_p(\mu, \nu) = \left(\min_{\pi \in \Pi_{\mu, \nu}} \int \|x - y\|^2 d\pi(x, y) \right)^{\frac{1}{p}}. \qquad (3.4)$$

3.2.3.4 Transporte Óptimo en adaptación de dominio

El TO se ha utilizado con éxito en el problema de adaptación del dominio. Algunos trabajos han utilizado el TO para encontrar un acoplamiento entre los dominios de origen y de destino, la idea es transportar los datos de origen y sus etiquetas al dominio de destino con un mapeo baricéntrico del acoplamiento. Finalmente realizaron el aprendizaje de un clasificador de las muestras transportadas en el dominio de destino. Se han estudiado una gran variedad de métodos basados en TO regularizado para crear conexiones óptimas entre muestras. Se ha propuesto una regularización de la norma de $l_p l_1$ para concentrar la información de transporte en elementos de la misma clase. Se ha utilizado también una regularización de lazo de grupo para promover la

transferencia masiva de datos de origen con las mismas etiquetas a datos de destino determinados. También algunas propuestas introducen una regularización laplaciana que tiene como objetivo preservar la estructura de datos aproximada por un gráfico durante el transporte. Por el lado teórico se justificó que la métrica de Wasserstein puede usarse como una medida de divergencia entre distribuciones para obtener garantías de generalización. Ahora en los enfoques más recientes, JDOT es un método en el que se utilizó el TO para encontrar un acoplamiento en las distribuciones conjuntas de datos y etiquetas. La función de costo del motor de la tierra incorporó un término en las características de los datos y en las etiquetas (Fatras et al., 2021).

3.2.4 Transporte Óptimo en redes neuronales profundas

3.2.4.1 *Aprendizaje supervisado y clasificación multi-etiqueta*

En aprendizaje supervisado, se quiere encontrar una función $f \in \mathcal{F}$ que describa la relación entre un vector aleatorio de características x y un vector objetivo aleatorio y , que siguen la medida conjunta $P(X, Y)$. Con este fin, se define una función de pérdida L que penaliza las diferencias entre la predicción $f(x)$ y el objetivo y . Luego, se define el promedio de la función de pérdida L sobre la medida de datos P , también conocida como el riesgo esperado (Fatras et al., 2021).

$$R(f) = \int L(f(x), y) dP(x, y). \quad (3.5)$$

La probabilidad $P(X, Y)$ conjunta se desconoce, aunque se tiene un acceso a las muestras empíricas extraídas de la probabilidad conjunta. Formalmente, sean los datos de una distribución idénticamente distribuida y sus etiquetas que se denotarán como $(x_1, y_1), \dots, (x_n, y_n)$, entonces, se aproxima así el riesgo esperado con la medida empírica. Luego se minimiza este riesgo conocido como minimización del riesgo empírico el cual se define como:

$$\arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n L(f_{\theta}(x_i), y_i), \quad (3.6)$$

donde f_{θ} es el clasificador parametrizado por un vector θ como una red neuronal profunda (Fatras et al., 2021).

3.2.4.2 *Clasificación multi-etiqueta y Transporte Óptimo*

En la clasificación de etiquetas múltiples, la etiqueta ya no es un vector de codificación unitaria y , por lo tanto, no es un vector de probabilidad. En este problema se codifica con varias informaciones en la etiqueta. Cada elemento de la etiqueta corresponde a una clase. Entonces, la red neuronal genera ahora un vector donde los elementos son todas las probabilidades de que las diferentes clases estén en la imagen. Para obtener una clasificación de probabilidad para cada elemento de salida, la última función de activación de una red neuronal suele ser una función sigmoidea. Sin embargo, la pérdida habitual L para la clasificación penaliza el error de forma isotrópica, es decir, dan la misma importancia a todos los errores entre todas las clases de forma independiente. La función de pérdida de transporte óptima multi-etiqueta se define de la siguiente manera. Sea $f_{\theta}: \mathcal{X} \mapsto (\mathcal{P})^{n_c}$ un mapa, donde \mathcal{P} denota una probabilidad y donde n_c es el número de clases. En la clasificación de etiquetas múltiples, la función $f_{\theta}(x)$ y la etiqueta y son medidas

de masa $\|f_\theta(x)\|_1$ y $\|y\|_1$ respectivamente. Por lo tanto, podemos usar el costo de TO desequilibrado como la función L para medir las diferencias entre la etiqueta y la salida del modelo, ya que las masas $\|f_\theta(x)\|_1$ y $\|y\|_1$ son generalmente diferentes. La fuerza de esta función de costo sobre las pérdidas de transporte óptimas es que puede tener en cuenta las similitudes de clase a diferencia de otras funciones de pérdidas. Las similitudes están codificadas en el costo C . Formalmente, el costo se puede calcular como $h(f_\theta(x), y, C)$, donde h es el costo de TO desequilibrado con regularización entrópica (Fratras et al., 2021).

3.2.4.3 Adaptación de dominio con Transporte Óptimo

En esta sección, se presenta el problema de la adaptación de dominio mediante TO, la cual consiste en transferir conocimiento de un conjunto de datos de origen a un conjunto de datos de destino utilizando los datos de origen etiquetados para clasificar los datos de destino sin etiquetar. La configuración de adaptación de dominio no supervisada es aquella en donde se tiene un conjunto de datos de dominio de origen etiquetado está dado por $D_S = \{(x_i^s, y_i^s)\}_{i=1}^n, x_i \in \mathbb{R}^d$, donde además se cuenta de un conjunto de datos de dominio de destino no etiquetado $D_t = \{(x_j^t)\}_{j=1}^n, x_j \in \mathbb{R}^d$. Existen otras variantes, como la adaptación de dominio semi-supervisada, donde están disponibles una cierta cantidad de etiquetas en el dominio de destino; además se tiene la adaptación de dominio supervisado, donde todas las etiquetas en el dominio de destino están disponibles. También hay extensiones de adaptación de dominio denominadas adaptación de dominio de múltiples fuentes o adaptación de dominio de múltiples objetivos, en las que tenemos acceso respectivamente a varios conjuntos de datos de origen o de destino. Se pueden también considerar otras variantes, como la adaptación de dominio parcial, donde las clases adicionales en el dominio de origen están presentes, pero no en el dominio de destino $y_S \subset y_t$; y la adaptación de dominio de conjunto abierto, donde las clases adicionales están en el dominio de destino, pero no en el dominio de origen $y_S \subset y_t$. La adaptación de dominio se puede presentar de dos formas:

- Clase desequilibrada: las distribuciones de etiquetas son diferentes en los dos dominios ($P_S(y) \neq P_t(y)$), pero las distribuciones condicionales de las muestras con respecto a las etiquetas son las mismas ($P_S(x^s|y) = P_t(x^t|y)$). Las configuraciones de adaptación de dominio de conjunto abierto y parcial son casos especiales de desequilibrio de clases.
- Cambio de covariable: las distribuciones condicionales de las etiquetas con respecto a los datos son iguales ($P_S(y|x^s) = P_t(y|x^t)$). Sin embargo, se supone que las distribuciones de datos en los dos dominios son diferentes ($P_S(x^s) \neq P_t(x^t)$).

Como podemos ver, y debido a su naturaleza el TO se puede utilizar con éxito en el problema de adaptación del dominio, además de su uso para encontrar un acoplamiento entre el origen y los dominios de destino (Fratras et al., 2021).

3.2.5 Adaptación de dominio con Transporte Óptimo en modelos de aprendizaje profundo

La adaptación de dominio toma en cuenta un dominio de origen que contiene datos emparejados $(X^s, Y^s) = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, donde $x_i^s \in \mathbb{R}^n$, $y_i^s \in \mathbb{R}^m$ este representa la entrada y la etiqueta asociada a cada muestra i . La adaptación de dominio, dado otro dominio al que se le denomina destino que contiene los datos no etiquetados $X^t = \{x_i^t \in \mathbb{R}^n\}_{i=1}^{N_s}$. El objetivo es estimar la

siguiente función tal que $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ para generar las etiquetas en el dominio de destino $Y^t = \{y_i^t\}_{i=1}^{N_t}$. La distribución de probabilidad de un conjunto de datos D , esta denotado por \mathbb{P}_D , donde D es X^s, Y^s, X^t y Y^t . El problema es que dada una función f tal que f induzca a una distribución de probabilidad $\mathbb{P}_{f(x^t)}$ en Y^t con $\mathbb{P}_{f(x^t)} \rightarrow \mathbb{P}_{Y^t}$ (Lin et al., 2021).

3.2.5.1 Modelo de Transporte Óptimo para adaptación de dominio en redes neuronales profundas

Dado un par de distribuciones \mathbb{P}_{D_1} y \mathbb{P}_{D_2} y una matriz de costos de desplazamiento $C \geq 0$, resolver con el Transporte Óptimo el plan de transporte $\gamma \in \Pi(\mathbb{P}_{D_1}, \mathbb{P}_{D_2})$ que minimiza el costo total.

$$\gamma \in \prod_{(\mathbb{P}_{D_1}, \mathbb{P}_{D_2})} \min \langle C, \gamma \rangle_F, \quad (3.7)$$

donde $\Pi(\mathbb{P}_{D_1}, \mathbb{P}_{D_2})$ denota el espacio de distribuciones marginales conjuntas \mathbb{P}_{D_1} ; y donde $\mathbb{P}_{D_1} \cdot \langle \cdot, \cdot \rangle_F$ es el producto de Frobenius, y la entrada C_{ij} de C representa el costo de desplazamiento de las i - ésima y j - ésima muestras (Lin et al., 2021).

3.2.5.2 Adaptación de domino por distribución conjunta

El mecanismo de adaptación se basa en la alineación entre las distribuciones conjuntas de los dominios de origen y de destino (para el dominio de destino, la etiqueta es una etiqueta estimada). Se aproxima f minimizando la pérdida de Transporte Óptimo entre las distribuciones conjuntas $\mathbb{P}_{X^s} \times \mathbb{P}_{Y^s}$ y $\mathbb{P}_{X^t} \times \mathbb{P}_{Y^t}$, con una matriz de costos elegida.

$$C_{ij} = \alpha \|x_i^s - x_j^t\|^2 + \beta \|y_i^s - f(x_j^t)\|^2, (\alpha, \beta) > 0. \quad (3.8)$$

Al alinear las distribuciones conjuntas de los dominios de origen y destino, la adaptación al ruido se logra de forma natural, ya que el Transporte Óptimo busca la muestra de origen que es más "similar" para cada muestra de destino (Damodaran et al., 2018; Lin et al., 2021).

3.2.5.3 Funciones de pérdida

La alineación de dominio de acuerdo con Damodaran et al., (2018) & Lin et al., (2021) se logra resolviendo el siguiente problema de optimización.

$$\min_{\gamma, f} \mathcal{L}_1 + \mathcal{L}_2 = \min_{\gamma, f} \frac{1}{N^s} \sum_i \|y_i^s - f(x_j^s)\|^2 + \sum_{ij} \gamma_{ij} (\alpha \|x_i^s - x_j^t\|^2 + \beta \|y_i^s - f(x_j^t)\|^2), \quad (3.9)$$

donde $\alpha, \beta > 0$ son los parámetros elegidos para el equilibrio.

3.2.5.4 Algoritmo para la adaptación de dominio usando el Transporte Óptimo

A continuación, se presenta el algoritmo general propuesto por los autores (Damodaran et al., 2018; Lin et al., 2021) para realizar la adaptación de dominio aplicando TO en redes neuronales profundas.

1. Se requieren las muestras del dominio fuente \mathbf{x}^s , las muestras de dominio objetivo \mathbf{x}^t , además de etiquetas del dominio fuente \mathbf{y}^s .
2. Para cada lote del dominio fuente ($\mathbf{x}^s, \mathbf{y}^s$) y del dominio objetivo (\mathbf{y}^t)
3. Ajustar θ_f , resolver γ de la ecuación
$$\min_{\gamma, f} \mathcal{L}_1 + \mathcal{L}_2 = \min_{\gamma, f} \frac{1}{N^s} \sum_i \|y_i^s - f(x_i^s)\|^2 + \sum_{i,j} \gamma_{ij} (\alpha \|x_i^s - x_j^t\|^2 + \beta \|y_i^s - f(x_j^t)\|^2)$$
por Transporte Óptimo.
4. Ajustar $\gamma, \theta_f \leftarrow \text{Adam}(\nabla_{\theta_f}, \mathcal{L}_2, \theta_f, \theta_h)$.

Este algoritmo general utiliza el enfoque del TO para adaptar los datos de origen al dominio de destino. El algoritmo ajusta los datos de origen al dominio de destino.

3.3 Modelos de aprendizaje profundo para regresión y clasificación

En esta sección se describen los fundamentos de las redes neuronales profundas a implementar en este trabajo de tesis. Este apartado aborda a las redes neuronales adversarias generativas (GANs, por sus siglas en inglés), las cuales se configuraron para resolver problemas de regresión y clasificación. Además, se detalla las redes neuronales que conforman a una GAN, como las redes neuronales convolucionales y las redes de memoria a largo plazo bidireccionales (BLSTM, por sus siglas en inglés).

3.3.1 Redes adversarias generativas

Las redes adversarias generativas son usadas en el aprendizaje semi-supervisado y no supervisado. Este tipo de redes neuronales son dos arquitecturas de redes neuronales que se entrenan en competencia simultánea (Creswell et al., 2018). Las redes adversarias generativas están compuestas por un generador y un discriminador que aprenden simultáneamente (ver figura 3.2). El generador aprende a generar muestras reales además de nuevas muestras de datos. El discriminador es un clasificador binario, que discrimina las muestras reales de las muestras generadas (Goodfellow et al., 2020). Las arquitecturas de redes que representan un generador y un discriminador se implementan mediante redes multicapa, que están compuestas por capas convolucionales y/o completamente conectadas. Las redes del generador y el discriminador deben ser diferenciables, aunque no es necesario que sean invertibles directamente (Creswell et al., 2018). El generador no tiene acceso directo a las muestras reales, su aprendizaje se produce a través de su interacción con el discriminador. El discriminador tiene acceso tanto a las muestras del generador además de contar con las muestras del conjunto de muestras reales. La propagación del error para el discriminador se proporciona con la información de saber si la muestra proviene del conjunto real o del generador. La misma señal de error, a través del discriminador, se puede utilizar para entrenar el generador, llevándolo a ser capaz de producir falsificaciones de mejor calidad. Si

consideramos que la red del generador mapea desde un espacio de representación, llamado espacio latente, hacia el espacio de los datos, podemos expresarlo de manera más formal como $G: G(z) \rightarrow R^{|x|}$, donde $z \in R^{|z|}$ es una muestra del espacio latente, $x \in R^{|x|}$ es una muestra, $|\cdot|$ y denota el número de dimensiones (Creswell et al., 2018). En una red adversaria generativa básica, la red del discriminador D se caracteriza de manera similar como una función que mapea los datos de imagen a una probabilidad de que la imagen provenga de la distribución de datos reales, en lugar de la distribución del generador: $D(x) \rightarrow [0,1]$. Para un generador fijo G , el discriminador D puede ser entrenado para clasificar imágenes como provenientes de los datos de entrenamiento (reales, cercano a uno) o del generador fijo (falso, cercano a cero). Cuando el discriminador es óptimo, se puede congelar o dejar de entrenar, y el generador G puede seguir siendo entrenado para reducir la precisión del discriminador. Si la distribución del generador logra coincidir perfectamente con la distribución de los datos reales, el discriminador estará máximamente confundido, prediciendo 0.5 para todas las entradas (K. Wang et al., 2017).

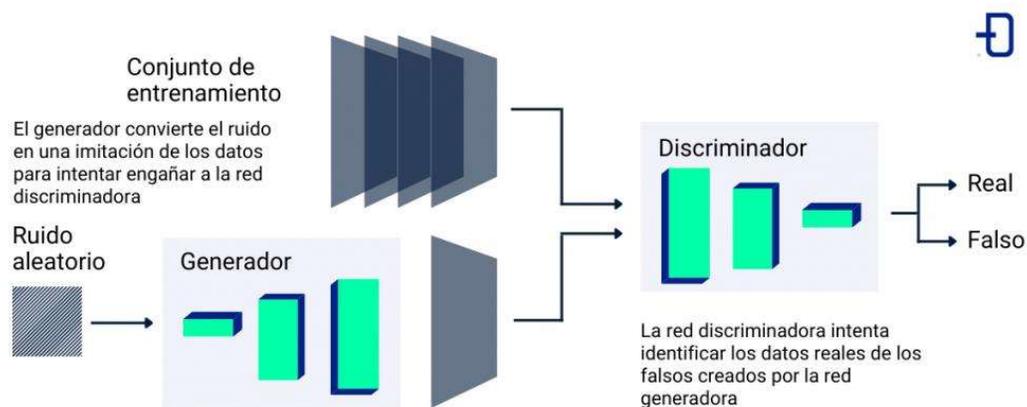


Figura 3.2 Red neuronal adversaria generativa

3.3.2 Red adversaria generativa como modelo de regresión

Las redes adversarias generativas diseñadas como modelo de regresión (ver figura 3.3) ayudan a producir muestras realistas que tengan propiedades específicas deseadas de las muestras. Este enfoque genera muestras con características específicas, este consta de un generador y un discriminador. En conjunto intentan producir muestras generadas con apariencia realista. El método consta de un generador, que es responsable de generar a partir de muestras con ruido el cual genera muestras realistas similares a las muestras en el conjunto de datos limpio, y un discriminador, que es responsable de validar las muestras generadas.

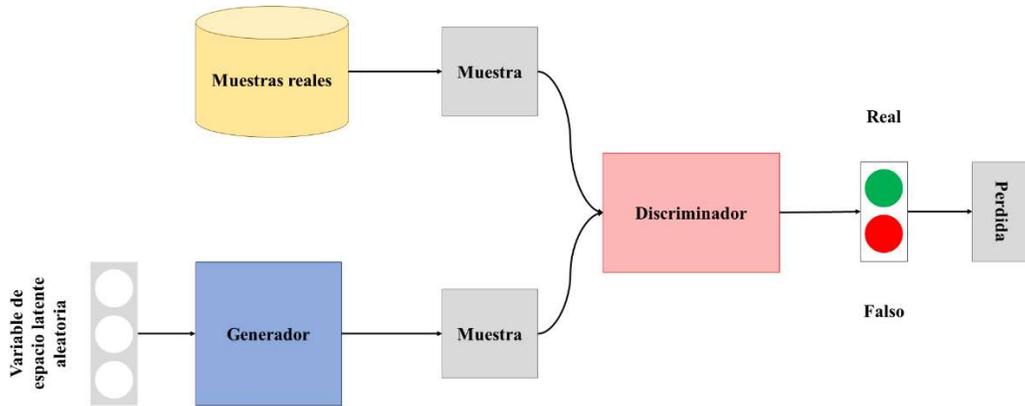


Figura 3.3 Red neuronal adversaria generativa como modelo de aprendizaje de regresión

3.3.3 Red adversaria generativa como clasificador

Las redes adversarias generativas diseñadas como modelo de clasificación (ver figura 3.4) constan de tres modelos separados: un generador, un discriminador y un clasificador. En cada iteración de entrenamiento, al generador se le proporcionan las muestras con ruido y genera muestras correspondientes. Luego, se actualiza el discriminador para mejorar su capacidad de distinguir entre muestras reales y generadas. Simultáneamente, se entrena un clasificador de manera estándar utilizando datos reales disponibles y sus respectivas etiquetas. Todos los datos reales disponibles tienen etiquetas en este método. Luego, se utilizan las muestras generadas como entradas para complementar la clasificación durante el entrenamiento (Haque, 2020).

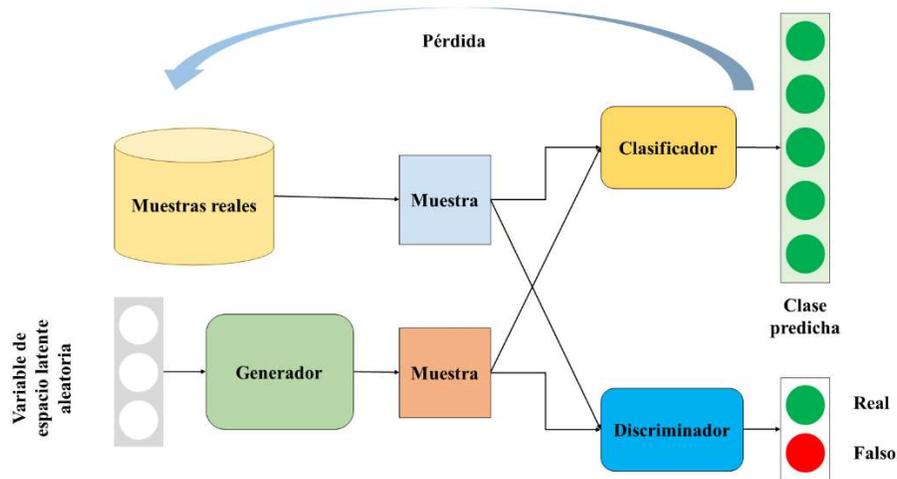


Figura 3.4 Red neuronal adversaria generativa - EC-GAN

3.3.4 Descripción de las arquitecturas de redes neuronales en las GANs

En este trabajo se describirán de manera general las arquitecturas de redes neuronales que

suelen conformar a las GAN implementadas. Estas GANs se utilizan como parte fundamental de la propuesta presentada en este proyecto.

3.3.4.1 Red neuronal convolucional

Las redes neuronales convolucionales están compuestas por tres tipos de capas. Estas son las capas convolucionales, las capas de agrupación y las capas completamente conectadas (ver figura 3.5). Cuando se apilan estas capas, se forma una arquitectura de red neuronal convolucional (O'Shea & Nash, 2015). La funcionalidad básica de la red neuronal convolucional se puede dividir en cuatro áreas clave:

1. Al igual que en otras formas de redes neuronales artificiales, la capa de entrada contendrá los valores.
2. La capa convolucional determinará la salida de las neuronas que están conectadas a regiones locales de la entrada a través del cálculo del producto escalar entre sus pesos y la región conectada al volumen de entrada. La unidad lineal rectificadora (ReLU) busca aplicar una función de activación, como la sigmoide o softmax, a la salida de la activación producida por la capa anterior.
3. La capa de agrupación realizará un submuestreo a lo largo de la dimensionalidad espacial de la entrada dada, reduciendo aún más el número de parámetros dentro de esa activación.
4. Las capas completamente conectadas realizarán las mismas funciones que se encuentran en las redes neuronales artificiales estándar e intentarán producir puntuaciones de clase a partir de las activaciones, que se utilizarán para la clasificación. También se sugiere que la función ReLU se pueda usar entre estas capas para mejorar el rendimiento.

Mediante este método sencillo de transformación, las redes neuronales convolucionales pueden transformar la capa de entrada original, capa por capa, utilizando técnicas de convolución y submuestreo para producir puntuaciones de clase para fines de clasificación y regresión.

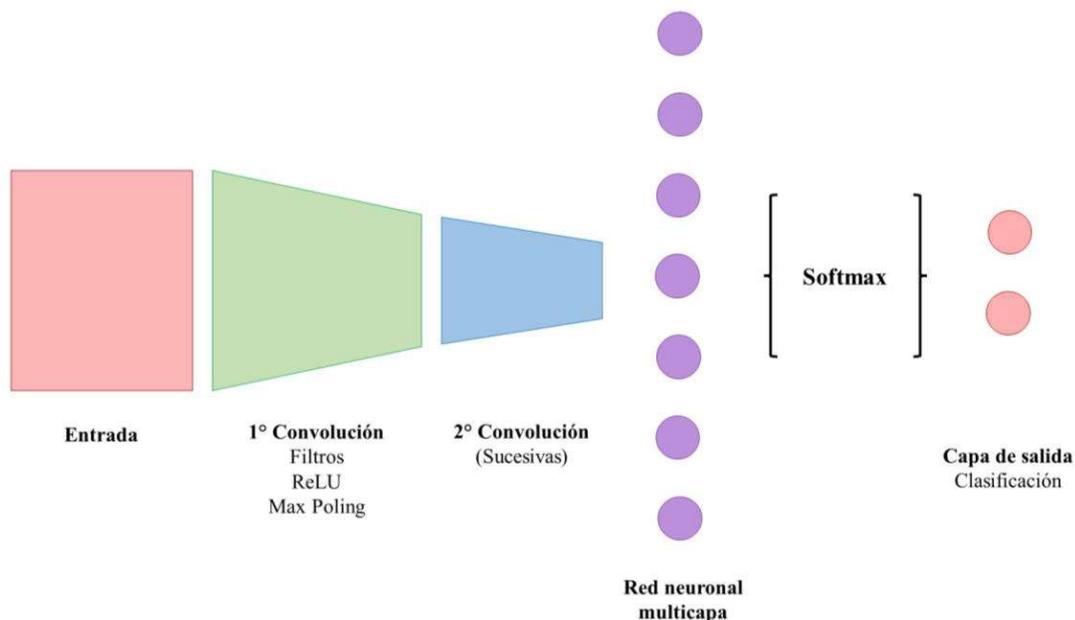


Figura 3.5 Red neuronal convolucional

3.3.4.2 VGG16

Una red neuronal convolucional VGG (ver figura 3.6) es un tipo de arquitectura de red neuronal profunda (Theckedath & Sedamkar, 2020). La red VGG-16 consta de 16 capas convolucionales y tiene un campo receptivo pequeño. Tiene una capa de agrupamiento de tamaño máximo y un total de 5 capas de este tipo. Después de la última capa de agrupamiento máximo, hay 3 capas completamente conectadas. Esto es seguido por tres capas completamente conectadas más. Utiliza el clasificador softmax como capa final. Aquí se aplica la función de activación ReLU a todas las capas ocultas (Rezende et al., 2018).

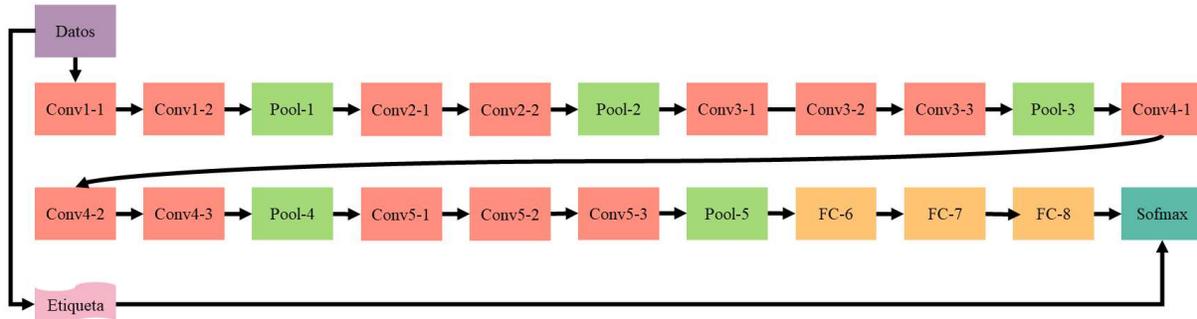


Figura 3.6 Arquitectura de red neuronal convolucional VGG-16

3.3.4.3 Redes neuronales recurrentes bidireccionales

Una red neuronal bidireccional de largo y corto plazo (BLSTM, por sus siglas en inglés) es un tipo de arquitectura de red neuronal recurrente utilizada para procesamiento de secuencias (Graves & Schmidhuber, 2005). La BLSTM es una variante de la red de Memoria a Largo Plazo y Corto Plazo (LSTM, por sus siglas en inglés), que tiene la capacidad de analizar la secuencia tanto en orden directo como en orden inverso (Graves et al., 2013). La idea básica de las BLSTM es presentar cada secuencia de entrenamiento hacia adelante y hacia atrás a dos redes recurrentes separadas, ambas conectadas a la misma capa de salida. Esto significa que, para cada punto en una secuencia dada, la red neuronal recurrente bidireccional tiene información completa y secuencial sobre todos los puntos anteriores y posteriores. Además, debido a que la red puede utilizar mucho o poco de este contexto como sea necesario, no se requiere encontrar un tamaño de ventana de tiempo (dependiente de la tarea) o un retraso objetivo (Huang et al., 2015). En la figura 3.7 se ilustra cómo las subredes hacia adelante y hacia atrás se combinan para llevar a cabo la tarea.

3.4 Transformada Wavelet

En esta sección se presenta la definición de la teoría base para la Transformada Wavelet, cuyo objetivo para este trabajo es su aplicación para la disminución de ruido en la señal de voz, antes de que la señal ingrese a los sistemas de reconocedores del habla.

3.4.1 Definición de Transformada Discreta Wavelet y sus propiedades

La transformada Wavelet utiliza funciones que están localizadas tanto en el espacio real como

en el de Fourier (Zhang, 2019). Esta transformada se puede expresar como:

$$F(a, b) = \int_{-\infty}^{\infty} f(x) \psi_{(a,b)}^*(x) dx, \quad (3.10)$$

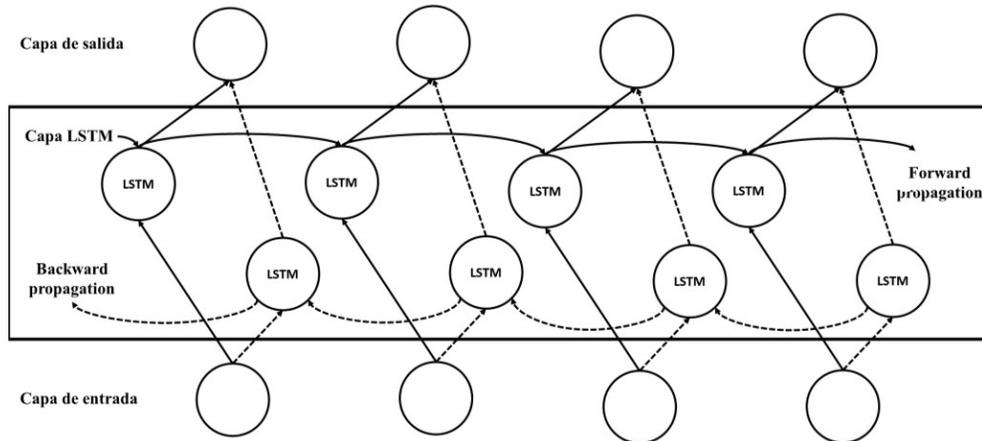


Figura 3.7 Red neuronal recurrente bidireccional

donde * es el símbolo complejo conjugado y la función ψ es alguna función base. Esta función puede elegirse arbitrariamente siempre que obedezca ciertas reglas. La transformada Wavelet es un conjunto infinito de varias transformadas, dependiendo de la función de mérito utilizada para su cálculo. También hay muchas formas de clasificar los tipos de transformadas Wavelets. Para este trabajo se abordará solo la división basada en la ortogonalidad de la Wavelet. Podemos usar Wavelets ortogonales para el desarrollo de transformadas discretas, que se aplicarán para este trabajo; y Wavelets no ortogonales para el desarrollo de transformadas continuas. Estas dos transformaciones tienen las siguientes propiedades:

- La transformada Wavelet discreta devuelve un vector de datos de la misma longitud que la entrada. Por lo general, incluso en este vector, muchos datos son casi cero. Esto corresponde al hecho de que se descompone en un conjunto de Wavelets (funciones) que son ortogonales a sus traslaciones y escalas. Por lo tanto, descomponemos dicha señal en un número igual o menor del espectro del coeficiente de Wavelet que el número de puntos de datos de la señal. Este espectro de Wavelets es muy bueno para el procesamiento y la compresión de señales.
- La transformada Wavelet continua, por el contrario, devuelve una matriz que es una dimensión más grande que los datos de entrada. Para un dato 1-D obtenemos una imagen del plano tiempo-frecuencia. Podemos ver fácilmente la evolución de las frecuencias de la señal durante la duración de la señal y comparar el espectro con otros espectros de señales. Como aquí se usa el conjunto no ortogonal de Wavelets, los datos están altamente correlacionados, por lo que aquí se observa una gran redundancia.

3.4.2 Transformada Wavelet Discreta

La transformada de Wavelet discreta es una implementación de la transformada que utiliza un

conjunto discreto de escalas y traslaciones de Wavelet que obedecen algunas reglas definidas (Zhang, 2019). En otras palabras, esta transformada descompone la señal en un conjunto ortogonal de Wavelets mutuamente ortogonales, que es la principal diferencia con la transformada continua, o su implementación para series de tiempo discretas. La Wavelet se puede construir a partir de una función de escala que describe sus propiedades de escala. La restricción de que las funciones de escala deben ser ortogonales a sus traslaciones discretas implica algunas condiciones matemáticas que se mencionan en la literatura especializada, la ecuación de dilatación se brinda a continuación:

$$\phi(x) = \sum_{k=-\infty}^{\infty} a_k \phi(Sx - k), \quad (3.11)$$

donde S es un factor de escala. Además, el área entre la función debe normalizarse y la función de escala debe ser ortogonal a sus traducciones enteras, es decir

$$\int_{-\infty}^{\infty} \phi(x)\phi(x + 1)dx = \delta_{0,1}. \quad (3.12)$$

Después de introducir algunas condiciones más se obtienen los resultados de todas estas ecuaciones, es decir, el conjunto finito de coeficientes a_k que definen la función de escala y también la Wavelet. La Wavelet se obtiene a partir de la función de escalado como N , donde N es un número entero par. El conjunto de Wavelets forma luego una base ortonormal que usamos para descomponer la señal. Por lo general, solo algunos de los coeficientes a_k son distintos de cero, lo que simplifica los cálculos.

3.4.3 Algoritmo para aplicar la técnica de filtrado basado en la transformada Wavelet

El algoritmo utilizado para mejorar la calidad de la señal se implementa mediante los siguientes pasos, descritos por Thu et al. (2019):

1. Se carga la señal de onda ruidosa.
2. A la señal de onda con ruido se le realiza una transformación logarítmica.
3. Se realiza una descomposición multinivel en la señal transformada logarítmica utilizando la transformada Wavelet.
4. Aplicar los tipos de Wavelet.
5. Aplicar el umbral a los coeficientes ruidosos.
6. Después de que los coeficientes de la señal descompuestos se umbralicen utilizando la técnica de umbralización, el sonido sin ruido se reconstruye utilizando la transformada Wavelet inversa.

3.4.3.1 Tipos de Wavelet

Existen diversos tipos de Wavelets que varían según las propiedades de las funciones Wavelets, como la ortogonalidad, la regularidad, la simetría y el soporte compacto. Algunos ejemplos son las funciones Haar, Daubechies, Symlets y Coiflets. La Wavelet Haar es la más simple, y por otro lado las Wavelets de Daubechies son muy populares. La transformada Wavelet tiene numerosas aplicaciones en las comunicaciones inalámbricas, como la flexibilidad en relación

con las portadoras, la sensibilidad a los canales y la reducción del consumo de energía al transferir datos mediante algoritmos de compresión de datos (Nyein Thu et al., 2008). En la figura 3.8 se muestran algunos de los tipos de Wavelets más comunes que se utilizan.

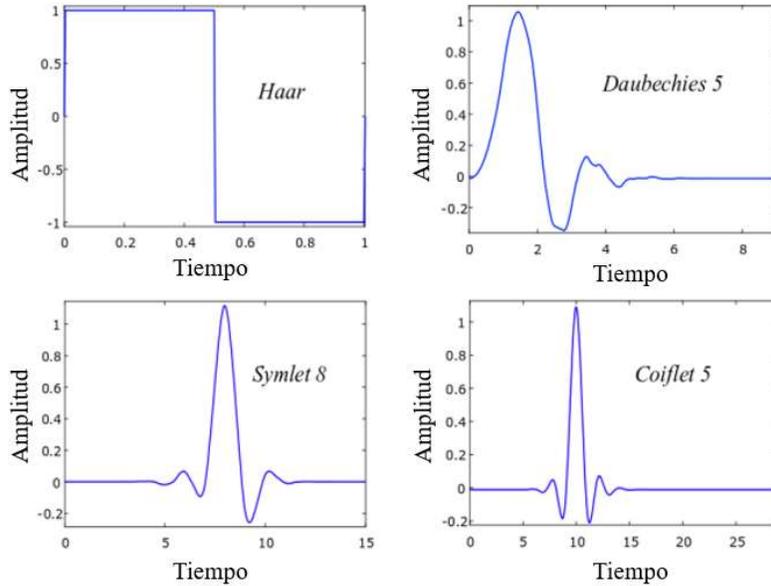


Figura 3.8 Ejemplos de Wavelets

3.4.4 Niveles de descomposición

La transformada Wavelet discreta analiza la señal de información en distintas bandas de frecuencia al descomponer la señal en diferentes coeficientes. Esta descomposición se logra mediante una serie de pasos sucesivos denominados pasa bajas y pasa altas a través de un filtro de indicación de dominio aislado (Parveen & Abdullah, 2019). La señal $x[n]$ se pasa entonces a través de $g[n]$, que es un filtro pasa altas; y $h[n]$ que es un filtro pasa bajas, como se muestra en la figura 3.9. Luego de filtrar, la señal se submuestra a la mitad. De acuerdo con la regla de Nyquist, la descomposición de la señal se expresa en términos de ecuaciones matemáticas de la siguiente manera:

$$y_n[k] = \sum_n x[n]g[2k - 1], \quad (3.13)$$

$$y_n[k] = \sum_n x[n]h[2k - 1]. \quad (3.14)$$

La descomposición duplica la resolución de la frecuencia. De manera similar, la señal se reconstruye mediante un par sucesivo de filtros de reconstrucción, como $g_1[n]$ y $h_1[n]$, que cumplen con la condición de reconstrucción perfecta.

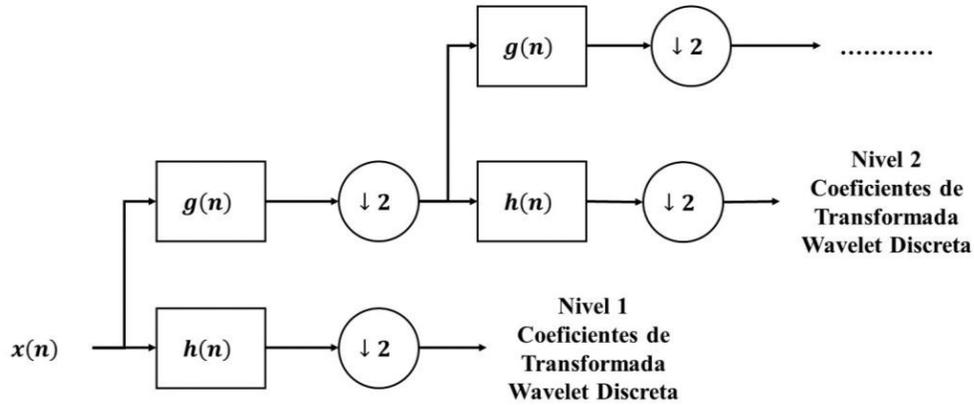


Figura 3.9 Niveles de descomposición de la Transformada Wavelet

3.4.5 Umbralización

Los métodos para la umbralización, de acuerdo con Thu et al., (2019), se descomponen en métodos duros (hard) y métodos suaves (soft), los cuales son los más populares y ampliamente utilizados para estimar los coeficientes para la eliminación de ruido mediante umbralización. La figura 3.10 muestra las señales de umbralización hard y soft respectivamente. La umbralización hard propone eliminar todos los valores por debajo de un umbral, mientras que los valores más altos se mantienen sin cambios. Una metodología alternativa podría utilizar, además de la cancelación, una resta del umbral a los valores restantes, que están por encima del umbral con el fin de reducir el número de discontinuidades en la señal.

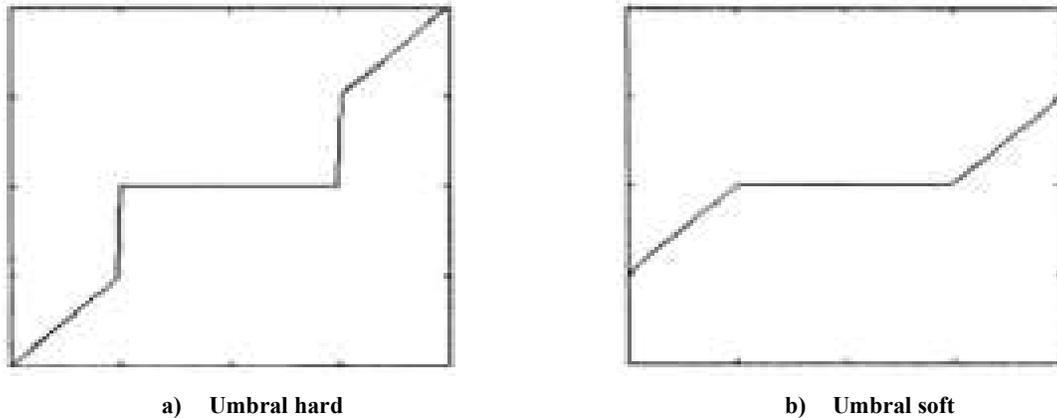


Figura 3.10 Tipos de umbralización

El principal problema en la eliminación de ruido, mediante umbralización, es la determinación del nivel de umbral adecuado.

3.5 Estudios relacionados

En esta sección se analizarán los estudios relevantes relacionados con la propuesta de tesis. En primer lugar, se abordan los estudios sobre los reconocedores automáticos de voz. A continuación, se discuten los trabajos relacionados con la identificación de locutores. Posteriormente, se examinan los estudios que aplican técnicas de adaptación de dominio en el reconocimiento del habla. Por último, se abordan las técnicas de filtrado y el aprendizaje profundo para mejorar el reconocimiento del habla en entornos ruidosos.

3.5.1 Estudios relacionados sobre los reconocedores de voz

En la literatura nos encontramos con varios enfoques para el caso de diseño de sistemas de reconocimiento automático de voz en ambientes con ruido, lo cual es importante mencionar para realizar la comparativa respecto a la propuesta realizada en este trabajo de tesis. Los autores Wang & Wang, (2016) propusieron una interfaz de separación de voz, basada en una red neuronal profunda y un modelo acústico basado en redes neuronales profundas, para construir una red neuronal más grande y ajustar conjuntamente los pesos en cada módulo. De esta manera, la interfaz de separación puede proporcionar el habla mejorada deseada por el modelo acústico y puede guiar a la interfaz de separación para producir una mejora más discriminatoria. El sistema logró una tasa de error de palabra promedio del 10.63 % en el conjunto de prueba del conjunto de datos sin ruido; mientras que, en la implementación con ruido, que representa el mejor rendimiento en este conjunto de datos, obtuvo una reducción de error del 22.75 %.

Por otro lado, Sun et al. (2019) propusieron un modelo robusto para aplicaciones en ambientes reales para los sistemas End-to-End, para ello formularon una regularización de una función objetivo con un entrenamiento adversario como ejemplos. Particularmente se centraron en el método del gradiente y el método de suavidad distribucional local. Su implementación permitió una mejora del 7%-12.2% de la reducción de tasa de error de palabra. Además, el trabajo de investigación de Tsao et al., (2017) propone la implementación de cuatro algoritmos basados en la restauración espectral de la señal de voz para el reconocimiento robusto; los cuales son un estimador espectral de error cuadrático medio mínimo, un estimador de amplitud espectral de máxima verosimilitud, un estimador de amplitud espectral a máximo posteriori y un algoritmo de amplitud espectral generalizado de máximo a posteriori. Donde la implementación con menor tasa de error de palabra corresponde al 5.35% y se obtuvo con el estimador de amplitud espectral de máxima verosimilitud.

3.5.2 Estudios relacionados sobre la identificación de locutor

La robustez de los sistemas de reconocimiento de locutores es crucial para aplicaciones del mundo real, las cuales suelen contener ruido. El trabajo para identificación de locutores en entornos ruidosos de X. Zhao et al., (2014) se basa en la eliminación del ruido de fondo usando mascarar binarias, donde utilizan como clasificador a las redes neuronales profundas. Este trabajo realiza una identificación de locutor robusta con modelos entrenados en condiciones de reverberación, que se basa en la marginalización acotada y una máscara directa. Sus resultados de evaluación muestran que el sistema propuesto mejora sustancialmente el rendimiento de la identificación de locutores en comparación con sistemas relacionados en una amplia gama de tiempos de reverberación y relaciones señal-ruido. Por otro lado, el trabajo de Ye & Yang (2021) menciona

que la mayoría de los métodos de investigación para la identificación de locutores actuales se basan en redes neuronales convolucionales o redes neuronales recurrentes. Los autores proponen un modelo de red neuronal profunda basado en una red neuronal convolucional bidimensional y una unidad recurrente con compuertas (GRU) para la identificación de locutores.

En estos trabajos se mencionan también que las características espaciales de la señal de voz, correspondientes al espectro de la voz, y las redes neuronales convolucionales son efectivas para la extracción de características en el caso de las características acústicas. Además, destacan que las redes neuronales recurrentes profundas, como las unidades recurrentes con compuertas (GRU), son mejores para representar enunciados largos en comparación con otro tipo de redes neuronales. Por otro lado, se ha intentado combinar características para la identificación de locutor, tal es el caso de Khdir et al. (2021), que utilizan nuevos métodos de extracción de características y una red neuronal de memoria a corto y largo plazo bidireccional para identificar al locutor. Ellos proponen combinar el espectrograma Mel y el cochleagrama para generar características más robustas y abundantes en enunciaciones cortas. Otros autores mantienen el esquema clásico, donde se mejora la identificación de locutor de entornos ruidosos. Ashar et al. (2020) utilizan una arquitectura novedosa con una red neuronal convolucional y coeficientes cepstrales de frecuencia de Mel (MFCC) para identificar al locutor en un entorno ruidoso. Se utiliza una técnica de extracción de características híbrida combinando la red neuronal convolucional y MFCC; donde lograron una precisión del 87.5% en la clasificación de locutores.

Enseguida, Bunrit et al. (2019) proponen un modelo de aprendizaje profundo utilizando una red neuronal convolucional para la identificación de locutor. Ellos utilizan un enfoque de texto independiente, donde cada 2 segundos de la voz del locutor se transforma en una imagen de espectrograma y se ingresa al modelo de una red neuronal convolucional entrenada desde cero. El método propuesto basado en red neuronal convolucional se compara con el método clásico de extracción de características utilizando coeficientes cepstrales de frecuencia de Mel (MFCC) y clasificación por máquina de soporte vectorial. Los experimentos realizados en habla tailandesa muestran que el método propuesto, basado en red neuronal convolucional entrenado en imágenes de espectrograma de voz, tiene un mejor rendimiento en comparación con los otros dos métodos, con una tasa promedio de clasificación del 95.83% para el conjunto de prueba. Los autores Chuang et al. (2019) proponen un sistema que combina un autoencoder de reducción de ruido profundo con una identidad de locutor incrustada para mejorar la calidad y la inteligibilidad de las señales de voz corrompidas por ruido aditivo. El sistema extrae características de identidad de locutor y utiliza el autoencoder para generar espectros mejorados. Ellos demuestran que el sistema propuesto logra mejoras significativas en la calidad del sonido y la inteligibilidad de la voz.

Los autores Qin et al. (2019) se centran en la verificación de locutor dependiente del texto en campo lejano. Utilizan un conjunto de datos pequeño y dependiente del texto en campo lejano junto con una base de datos grande y sin dependencia del texto en habla cercana para el entrenamiento. Demuestran que los datos independientes del texto en campo lejano y el ajuste fino del modelo de incrustación de locutor pre-entrenado pueden mejorar significativamente el rendimiento del sistema. Ellos agregaron ruidos reverberantes a los datos de inscripción limpios para mejorar aún más el rendimiento del sistema en aplicaciones reales. Finalmente, el trabajo propuesto por F. Zhao et al. (2019) lleva a cabo un marco de verificación de locutores de tipo End-to-End para mejorar la robustez frente al ruido de fondo. El marco utiliza una red convolucional recurrente para abordar la separación del habla y se optimiza en conjunto con el sistema de verificación de locutores. Se utiliza la salida de la capa intermedia de la red convolucional recurrente como característica auxiliar junto con la característica robusta de bancos de filtros del

habla ruidosa. Ellos muestran que el algoritmo propuesto tiene un mejor rendimiento en condiciones ruidosas.

3.5.3 Estudios relacionados de adaptación de dominio en el reconocimiento voz

Como se planteó al inicio del presente trabajo, esta investigación busca realizar una adaptación de dominio para la disminución y/o eliminación de discrepancias que puedan degradar las tasas de rendimiento en el reconocimiento de voz. Un trabajo enfocado a este tipo de implementación, donde se aplicó aprendizaje no supervisado y la implementación de adaptación de dominio para reconocimiento de voz, es el propuesto por Khurana et al. (2020). Este trabajo menciona que, dentro del sistema de reconocimiento de voz, el rendimiento se degrada significativamente cuando los dominios de datos de entrenamiento y prueba no coinciden debido a los diferentes factores de incertidumbre que afectan al reconocedor de voz. Este trabajo propuso el auto entrenamiento mediante redes neuronales profundas, donde combinaron un enfoque de filtrado de pseudo-etiquetas basado en la incertidumbre donde se utilizaron de manera efectiva para la adaptación del dominio. La propuesta excluye los datos pseudoetiquetados con incertidumbres altas del entrenamiento. La propuesta de Khurana et al. (2020) menciona que alcanzo un 80% del rendimiento de un sistema entrenado con datos reales.

Según Hwang et al. (2022), utilizaron la combinación de métodos de aprendizaje auto-supervisados y semi-supervisados para resolver problemas de adaptación de dominios en un entorno de producción a gran escala para modelos del reconocimiento del habla. Ellos implementaron redes neuronales recurrentes con transductor. Su enfoque demuestra que el uso de los datos del dominio de origen con una pequeña fracción de los datos del dominio de destino (3%) puede recuperar la brecha de rendimiento en comparación con una línea base de datos completa, alcanzando 13.5% de mejora relativa de la tasa de error de palabra para los datos del dominio de destino. Por otro lado, Ali et al. (2022) implementaron la extracción de características por coeficientes cepstrales de frecuencia de Mel (MFCC). Se menciona que utiliza una red neuronal de codificador automático dispersa para clasificar, mientras que se usa un modelo oculto de Markov para decidir sobre el reconocimiento de voz. El rendimiento de la red está optimizado, ya que aplico el algoritmo de optimización Harris Hawks para ajustar los parámetros de la red. La red afinada puede reconocer efectivamente el habla en un entorno ruidoso, la propuesta reconoce la síntesis de voz con un 99.31% de precisión, un 99.22 % de recall, un 99.21 % de Coeficiente de correlación de Matthew y un 99.18 % de valor de F-Measure.

Los autores Sokolov & Savchenko (2021) se centraron en la puesta a punto de modelos acústicos para objetivos de adaptación de genero de locutor. Ellos entrenaron previamente modelos basados en Transformers, además llevaron a cabo mediante ajustes finos en los subconjuntos de prueba específicos de género para mejorar el modelo. Los resultados que obtuvieron respecto a la tasa de error de palabra obtenida con respecto a la línea de base son de un 5%, mientras que un 3% más baja en los subconjuntos masculino y femenino, respectivamente. Esto es alcanzado siempre y cuando las capas del codificador y del decodificador no están congeladas y la sintonización se inicia desde los últimos puntos de control. Vinculado a esto, los autores Sim et al. (2018) mencionan que la solidez del dominio es un problema desafiante para el reconocimiento automático de voz. En su trabajo consideraron los datos de voz recopilados para una capa oculta factorizada de la red neuronal, el cual trataron como una representación compacta de rango bajo para adaptar un sistema reconocedor automático de voz multidominio a dominios invisibles. Los

resultados experimentales en dos dominios invisibles muestran que la capa oculta factorizada es un método de adaptación más efectivo en comparación con el ajuste fino selectivo de parte de la red. El rendimiento de los modelos ajustados mejora aproximadamente de forma lineal con un número creciente de capas ajustadas. El ajuste fino de las 4 capas de una red neuronal de Memoria a Largo y Corto Plazo (23.6 millones de parámetros) logra una tasa de error de palabra del 10.8 %. Por otro lado, la adaptación de la capa oculta factorizada con el rango 20 logró una tasa de error de palabra del 1.6% utilizando solo 0.5 millones de parámetros específicos del dominio. Asimismo, el trabajo de Ahn et al. (2021) implementó redes neuronales profundas para el reconocimiento de emociones del habla en un escenario de corpus cruzado. Ellos propusieron un reconocimiento de emociones del habla entre corpus basado en el aprendizaje y la adaptación del dominio no supervisado, que está entrenado para aprender la similitud de clase (emoción) de las muestras del dominio de origen y lo adapta al dominio de destino. Utilizaron múltiples corpus en el entrenamiento para mejorar la solidez del reconocimiento de emociones en las muestras invisibles. Los experimentos en corpus de habla emocional con tres idiomas diferentes mostraron que el método propuesto superó a otros enfoques. Su propuesta aprende a comparar las características de la consulta con las del conjunto de soporte para las clases.

Wei-Ning et al. (2017) realizaron un entrenamiento con un codificador automático variacional en datos de dominio de origen y de destino (sin supervisión) para aprender una representación latente del habla. Luego, transformaron los atributos molestos del habla que son irrelevantes para el reconocimiento modificando las representaciones latentes, esto para aumentar los datos de entrenamiento etiquetados con datos adicionales cuya distribución es más similar al dominio objetivo. Ellos reportaron una reducción de la tasa absoluta de error de palabra hasta en un 35% en comparación con la línea de base no adaptada. Sun et al. (2017) abordan el problema del reconocimiento de voz robusto como una tarea de adaptación del dominio. Específicamente se enfocaron en la adaptación de dominio profundo no supervisado para el modelado acústico con el fin de eliminar el desajuste entre la prueba y el entrenamiento, que es común en el uso del reconocimiento de voz en el mundo real. Bajo un marco de aprendizaje multitarea, el enfoque aprende conjuntamente dos clasificadores discriminativos utilizando una red neuronal profunda. Como tarea principal, un predictor de etiquetas predice etiquetas de fonemas y se usa durante el entrenamiento y en el momento de la prueba. Como segunda tarea, un clasificador de dominios discrimina entre los dominios de origen y de destino durante el entrenamiento. La red se optimiza minimizando la función de pérdida del clasificador de etiquetas y maximizando la función de pérdida del clasificador de dominio al mismo tiempo. Su enfoque no supervisado solo necesita datos de entrenamiento etiquetados del dominio de origen y algunos datos en bruto no etiquetados del nuevo dominio. Los experimentos que llevaron a cabo con el reconocimiento de voz sobre ruido/distorsión de canal y cambio de dominio confirman la eficacia del enfoque propuesto. El enfoque logra una reducción relativa de la tasa de error de palabra del 37.8%.

Finalmente, el trabajo de Bell et al. (2021) plantea por otro lado, una descripción general estructurada de los algoritmos de adaptación para el reconocimiento de voz basado en redes neuronales, considerando tanto los modelos ocultos de Markov y redes neuronales con un enfoque en la adaptación del hablante, la adaptación del dominio y la adaptación del acento. La descripción general caracteriza los algoritmos de adaptación basados en incrustaciones, adaptación de parámetros del modelo o aumento de datos.

3.5.4 Estudios relacionados sobre la mejora del habla con transformada Wavelet

Los principales estudios relacionados sobre la aplicación de la Transformada Wavelet son los que se comentan a continuación. En esta sección se discutirán sus principales aportes en el reconocimiento de voz y cómo su aplicación influye para aumentar el rendimiento de un sistema de RAV. El trabajo de Soe Naing et al. (2020) implementó en un sistema de RAV la transformada discreta Wavelet para la eliminación de ruido. Su propuesta fue abordada en las arquitecturas de reconocimiento de voz, que son modelos de mezclas Gaussianas con modelos ocultos de Markov y redes neuronales profundas con modelos ocultos de Markov. El rendimiento de su desarrollo mostró que las características resistentes al ruido, tales como en un subterráneo de metro, balbuceo, coches y conversaciones, se obtienen mientras se combinan con la eliminación de estos ruidos mediante la transformada de Wavelet de los coeficientes ceptrales de frecuencia de Mel (MFCC). La mejor precisión que obtuvieron fue con el entrenamiento de DNN-HMM de entropía cruzada mediante la eliminación de ruido con Wavelets Coiflet y el umbral Rigrsure, que proporciona un 97.54 % en 10 dB, un 93.13 % en 5 dB, un 75.63 % en 0 dB y un 37.29 % en -5 dB. A diferencia del trabajo anterior, donde aplica la transformada Wavelet a todo el proceso MFCC, la investigación de Hidayat & Winursito (2020) se centró en el análisis de la etapa de transformada rápida de Fourier de los MFCC. El método propuesto realizó el proceso de eliminación de ruido utilizando la transformada Wavelet solo en los datos relacionados con el ruido en función de los resultados del análisis del proceso de la transformada rápida de Fourier. El estudio utilizó datos del habla en forma de once palabras aisladas en inglés, a las que se añadió ruido con varias características diferentes. Los resultados mostraron que el método propuesto era capaz de generar una mayor precisión que los métodos convencionales de eliminación de ruido; esto se dio usando una relación señal/ruido de 10dB, 15dB y 20dB utilizando la onda Fejer Korovkin 6. El mayor aumento de precisión del método que propusieron fue en la relación señal/ruido (SNR) de 15 dB con un aumento del 4.63%, seguido de un aumento del 3.96% a una intensidad de 20 dB y un 2.3% a una intensidad de 10 dB.

Los autores Ashwin & Manoharan (2018) propusieron un método de umbral de bloque en la transformada de Fourier en tiempo corto, donde se utilizó para eliminar el ruido de la señal de audio de manera efectiva. Por otro lado, Abdullah et al. (2021), mediante la transformada Wavelet discreta, proponen un detector de actividad de voz adaptativo y una selección de sub-bandas para la extracción de características para la clasificación de ruido, que se puede utilizar en una canalización de procesamiento de voz. En comparación con la técnica convencional basada en la transformada de Fourier de tiempo corto, tiene puntajes F1 y precisiones de clasificación más altas (con una media de 0.916 y 90.1%, respectivamente) en cinco tipos de ruido diferentes (balbuceo, fábrica, automóvil, ruido blanco, y ruido rosa).

3.5.5 Estudios relacionados sobre la mejora del habla con aprendizaje profundo

Algunos enfoques para la mejora del habla ruidosa se basan en el aprendizaje profundo, el cual ha tenido un impacto para mejorar la precisión en este tipo de sistemas, es por ello por lo que se han estudiado las redes neuronales profundas para mejora del habla, tal fue el trabajo de Nossier et al. (2021), que proporciona un estudio de siete arquitecturas de redes neuronales profundas de tipo perceptrón multicapa, red neuronal convolucional y codificador-autoencoder. Por otro lado,

es importante como se trata la mejora del habla en cuanto a la caracterización de la señal de voz para los modelos de aprendizaje profundo. Para ello, algunos trabajos se han centrado en estimar la magnitud limpia de la transformada de Fourier con redes neuronales profundas, tal como lo mencionan Liao et al. (2019), Bhat et al. (2019) y Tan et al. (2019). Otros trabajos, como el de Choi et al. (2019) y Hu et al., (2020), proponen bloques de construcción para manejar espectrogramas de valores complejo a través del aprendizaje profundo para mejorar la señal de voz. También se ha tratado de mejorar el habla sin aplicar ningún procesamiento a la señal de voz, tal es el caso de Defossez et al. (2020), que utilizan la forma de onda de la señal de voz. Dado el gran aporte que se tiene respecto a la estimación de la magnitud limpia del espectro de Fourier este trabajo se centra en este sentido.

Ahora, pasando a analizar las distintas arquitecturas de redes neuronales profundas, se tiene que las redes neuronales generativas han mostrado tener una mayor eficacia en estimar el habla limpia, tal como proponen los autores Phan et al. (2020), G. Park et al. (2020) y Donahue et al. (2018); donde ellos evalúan la efectividad de las redes neuronales generativas para la mejora del habla. Es por ello por lo que el trabajo propuesto en esta tesis se centra en usar este tipo de redes neuronales. Los trabajos anteriores se centran solo en mejorar el habla en un solo dominio, ahora analizaremos los trabajos hechos respecto a la adaptación de dominio del modelo de aprendizaje profundo. La adaptación de dominio se ha desarrollado en conjunto con los enfoques de las redes neuronales generativas. Se tiene que Zhao et al. (2017) se centran en múltiples dominios empleando varios generadores en las redes adversarias generativas; mientras que Y. Li et al. (2022) abordan el problema de un desajuste de tipo de ruido entre las condiciones de entrenamiento y prueba. Ambos enfoques se centran en la adaptación de dominio con entrenamiento adversario. Por otro lado, se ha estudiado la adaptación de dominio no supervisado, donde solo se necesitan las muestras de audio limpio en el dominio destino. Finalmente, el trabajo de Y. Li et al. (2022) propone un método novedoso de adaptación de dominio empleando el Transporte Óptimo, donde se propone la mejora del habla no supervisada. Dado que nuestro trabajo pretende combinar técnicas de aprendizaje profundo con técnicas de filtrado basadas en Wavelets, tenemos que Meriane Brahim (2021), Naing et al. (2020), Patil (2015), Abdullah et al. (2021a) y Chelali et al. (2018) proponen usar filtrado mediante transformada Wavelet. En sus trabajos demostraron eliminar el ruido de forma eficiente en diferentes niveles de relación señal-ruido.

Capítulo 4 Método y propuesta de investigación

En el presente capítulo se llevará a cabo un análisis del método de investigación empleado en este trabajo, junto con la exposición de la propuesta de investigación desarrollada. Se abordan aspectos fundamentales, como la configuración del problema y las configuraciones adecuadas que se implementaron para llevar a cabo el proceso experimental.

4.1 Modelo de investigación

La metodología empleada en este trabajo se compone de seis etapas, tal como se muestra en la figura 4.1. Esta metodología se basa en el esquema general de investigación presentado por Hernández et al. (2017), que describe el modelo científico. De acuerdo con lo representado en la figura 4.1, la primera etapa se enfoca en la formulación del problema abordado en esta investigación, que consiste en mejorar el reconocimiento del habla en entornos ruidosos. La segunda etapa se centra en realizar una revisión crítica de la literatura, que comprende la exploración de modelos de arquitecturas para la eliminación y/o reducción de ruido en la señal de voz. En esta etapa se analizan detalladamente las características más eficaces utilizadas en la eliminación de ruido en el habla, así como las arquitecturas de redes neuronales que han demostrado mayor efectividad en esta tarea. Además, se examinan las técnicas de filtrado relevantes para abordar la tarea de eliminación de ruido. La tercera etapa tiene como objetivo la selección de las técnicas de procesamiento más adecuadas para el reconocimiento del habla en entornos ruidosos, basándose en la efectividad reportada en cada uno de los experimentos según el estado del arte. En la cuarta etapa, se presenta la arquitectura diseñada para mejorar el reconocimiento del habla en entornos ruidosos. Además de seleccionar las técnicas apropiadas, es importante definir el orden de aplicación de acoplamiento para la eliminación y/o reducción de ruido, dado que, según lo expuesto en el estado del arte, la aplicación de un segundo procesamiento en la señal de voz podría ocasionar una pérdida de información. La quinta etapa abarca el desarrollo conceptual y experimental de la propuesta de investigación para eliminar y/o reducir el ruido en diversos entornos ruidosos. Finalmente, en la etapa 6 se valida el modelo de mejora del habla utilizando las métricas de STOI y PESQ. Asimismo, se evalúa el rendimiento del habla mejorada mediante el reconocimiento automático de voz y la identificación de locutores, independientemente del contexto.

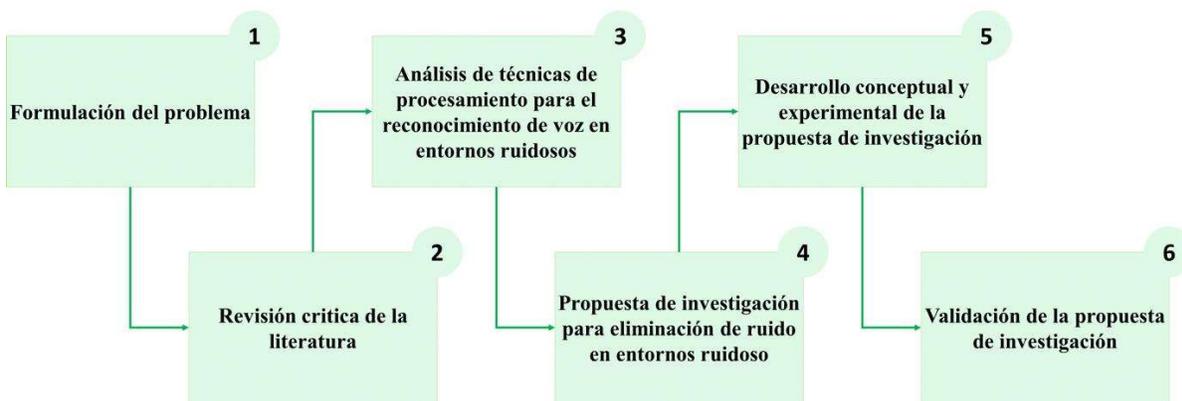


Figura 4.1 Modelo de investigación general

4.2 Propuesta de trabajo a realizar

El trabajo realizado se basa en el flujo general de la metodología propuesta, como se muestra en la figura 4.2. Esta metodología se centra en mejorar el reconocimiento del habla mediante la implementación de técnicas de filtrado y aprendizaje profundo, con el uso de adaptación de dominio como parte del proceso para reducir y/o eliminar el ruido presente en la señal de voz. El propósito de combinar estos dos enfoques es lograr un reconocimiento del habla más robusto frente al ruido, que contribuya a obtener resultados más precisos y reducir las incidencias erróneas en los sistemas de reconocimiento del habla.

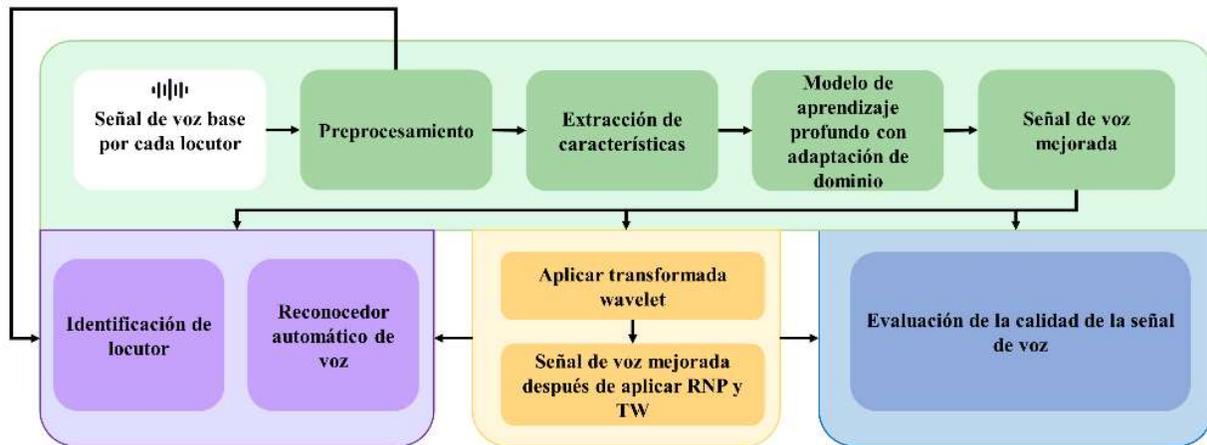


Figura 4.2 Propuesta de trabajo para el sistema de reconocimiento del habla en entornos ruidosos

4.2.1 Descripción de la propuesta de trabajo

A continuación, se presenta en detalle la propuesta de trabajo para desarrollar un modelo robusto de eliminación de ruido en el reconocimiento del habla en entornos ruidosos.

1. El primer bloque se centra en tomar como entrada una señal de audio digital para aplicarle un preprocesamiento a la señal de audio. En este sentido, se asegura que los conjuntos de audio de habla y ruido estén en un solo canal y tengan la misma frecuencia de muestreo. En este mismo bloque se generan conjuntos de audios base y con incrustación de ruido, y se divide el conjunto de datos en los dominios de origen y objetivo. Cabe señalar que los diferentes tipos de ruido fueron añadidos a las señales de voz de forma aleatoria, conservando una distribución uniforme en cuanto al ruido añadido y los niveles SNR aplicados. Durante la extracción de características, la señal es convertida del dominio del tiempo al dominio de las frecuencias y se extraen los vectores de características adecuados. Durante esta fase de extracción de características se extrae la magnitud del espectro de Fourier de los conjuntos de datos para luego someterlos al modelo de aprendizaje profundo. La fase de la transformada de Fourier se conserva para aplicar la transformada inversa de Fourier y obtener el audio del habla mejorado.
2. Posteriormente, en el segundo bloque los audios mejorados por la red neuronal con adaptación de dominio se someten a una evaluación donde se calculan las métricas de STOI y PESQ.

3. Los audios obtenidos pasan al tercer bloque, donde se aplica la eliminación de ruido basado en la transformada Wavelet. El resultado de este bloque es evaluado en el segundo bloque.
4. Finalmente, los resultados obtenidos por la red neuronal con adaptación de dominio y la transformada Wavelet son sometidos a los sistemas de reconocimiento del habla, que consiste en integrar un modelo acústico, y este tendrá como parámetros la información acústica y fonética; además, se toma como entrada el resultado de la extracción de características, y se generan las probabilidades del modelo acústico. El cálculo de parámetros en esta etapa se realiza mediante redes neuronales profundas, los cuales se encuentran en el bloque 4. En este bloque, se utiliza un modelo pre-entrenado para el reconocimiento automático de voz y se entrena un modelo de reconocimiento de locutor donde posteriormente es evaluado a través de las métricas de clasificación.

4.3 Selección y preparación de los conjuntos de datos

En esta sección se describen en detalle los conjuntos de audios utilizados para el reconocimiento del habla en entornos ruidosos, así como el proceso de preparación llevado a cabo en dichos conjuntos para la ejecución de la propuesta de trabajo.

4.3.1 Corpus para el reconocimiento del habla

En este trabajo experimental se utilizó el conjunto de datos de Panayotov et al. (2015) de LibriSpeech. Este es un corpus de grabaciones de lecturas en inglés de libros de dominio público. Contiene grabaciones de hablantes con distintos estilos de habla, totalizando alrededor de 1000 horas de audio. Este incluye segmentos de audio de 10-30 segundos, junto con transcripciones de texto para cada grabación. Para la eliminación de ruido en la señal de voz se seleccionaron 20 locutores del conjunto de datos, cada uno con 100 audios. Este subconjunto se dividió en 80% para la fase de entrenamiento y 20% para la fase de validación. Para la fase de pruebas del reconocimiento del habla, se procedió a seleccionar un subconjunto de 10 locutores del conjunto de datos original. A partir de estos locutores, se tomaron aleatoriamente 100 audios, a los cuales se les aplicaron los respectivos procesamientos. Estos audios fueron utilizados para evaluar los resultados en un RAV pre-entrenado de Watanabe (2021). En cuanto a la fase de identificación del locutor, se emplearon los mismos 10 locutores seleccionados anteriormente. Se conservaron los audios del subconjunto generado en la fase de pruebas. Luego se dividió el conjunto de datos en un 80% para el entrenamiento y un 20% para la validación del modelo.

4.3.2 Corpus con ruido natural

El conjunto de datos de audios utilizados para añadir ruido a la señal de voz del locutor proviene de varias fuentes. Para los ruidos de lluvia, helicóptero, motor, campanas de iglesia, llanto de bebé, gotas de agua y escritura en teclado, se utilizó el conjunto de datos Dataset for Environmental Sound Classification de Piczak (2015). Para los ruidos como ladrido de perro, música de fiesta y aire acondicionado, se utilizó el conjunto de datos UrbanSound8K de Salamon et al. (2014). Además, se utilizó el conjunto de datos DEMAND de Thiemann et al. (2013a) para los ruidos de cafetería, murmullo de personas y ruido de coche. También se generó ruido rosa utilizando Python. En total, se trabajó con 14 tipos de ruido en el estudio, lo cual se describe en

detalle en la tabla 4.1 y cómo se dividió el ruido en fuente y objetivo para la aplicación de la adaptación de dominio.

Tabla 4.1 Conjuntos de ruido

Ruido fuente	Ruido objetivo
1. Ruido rosa	1. Llanto de bebe
2. Gotas de agua	2. Multitud de gente
3. Carro	3. Campanas de iglesia
4. Cabina	4. Mormullos de cafetería
5. Lluvia	5. Helicóptero
6. Viento	6. Personas hablando
7. Escritura de teclado	7. Ladridos de perro

4.3.3 Generación del conjunto de habla ruidosa

En la figura 4.3 se muestra el proceso para añadir ruido a la señal de voz con base a la relación señal-ruido. Como entrada es necesario ingresar la señal del habla base, ésta entra a través de un canal de distorsión donde se le añade ruido, tanto fuente como ruido objetivo de la tabla 4.1 en los diferentes niveles de SNR [-9, -6, -3, 0, 3, 6, 9] dB.

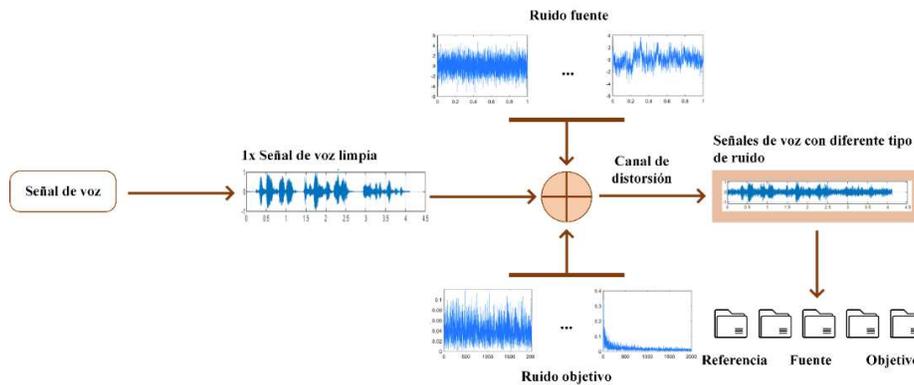


Figura 4.3 Generación del conjunto del habla ruidosa

4.3.3.1 Proceso para añadir ruido a la señal de voz con base a la relación señal-ruido

1. Dada la muestra $s(n)$ de la señal de voz en el instante n , se calcula la potencia de la señal voz.

$$\text{Potencia de la señal de voz} = \frac{1}{N} \sum_{n=1}^N s(n)^2, \quad (4.1)$$

donde N es el total de muestras.

2. La potencia de la señal está dada por la ecuación:

$$\text{Potencia de la señal de voz (dB)} = 10 \log_{10}(\text{Potencia de la señal de voz}). \quad (4.2)$$

3. Se calcula la potencia de la señal de ruido con la ecuación:

$$\text{Potencia de la señal de ruido} = \frac{\text{Potencia de la señal de voz (dB)}}{10^{\frac{SNR}{10}}}, \quad (4.3)$$

donde SNR corresponde a la relación señal-ruido. Además, entre menor sea el SNR, la amplitud del ruido es mayor que la señal de voz, y entre mayor sea la SNR, la amplitud del ruido es menor que la de la señal de voz.

4. Calcular la potencia del ruido con la ecuación:

$$\text{Potencia del ruido (dB)} = 10 \log_{10}(\text{Potencia de la señal de ruido}). \quad (4.4)$$

5. Se calcula el nivel SNR para añadir a la señal de voz el ruido.

$$\text{Señal ruidosa} = \sqrt{\text{Potencia de la señal de ruido}} \frac{v(n)}{\sigma(v(n))}. \quad (4.5)$$

6. Finalmente se tiene la ecuación resultante que genera el habla ruidosa.

$$\text{Habla ruidosa} = \text{señal de voz} + \text{señal ruidosa}. \quad (4.6)$$

4.4 Configuración del problema del habla ruidosa

El problema de reconocimiento del habla en entornos ruidosos plantea un desafío debido a la naturaleza no estacionaria de la señal de voz. Una solución ampliamente utilizada para abordar este problema es la transformada de Fourier de tiempo corto, la cual permite convertir la señal de voz en una señal cuasi-estacionaria. La transformada de Fourier de tiempo corto proporciona información precisa sobre las frecuencias presentes en la señal en diferentes momentos, lo cual resulta especialmente útil cuando los componentes de frecuencia varían a lo largo del tiempo (Kehtarnavaz, 2008).

$$X(\omega, m) = \sum_{k=0}^{K-1} w(k) \cdot x(m \cdot H + k) \cdot \exp\left(-j \frac{\pi \omega k}{K}\right), \quad (4.7)$$

donde:

n_fft : tamaño de la transformada de Fourier

m : es el índice de la ventana

ω : es la frecuencia $0 \leq \omega \leq n_fft$

K : tamaño de ventana

k : índice de ventana

$w(k)$: segmento de ventana

x : señal de voz en 1-D

H : espaciado entre ventanas

De acuerdo con Bhat et al. (2019), Boyko & Hrynyshyn (2021) y Roy et al. (2021), las técnicas

de la mejora del habla $y(n)$ considera una mezcla de señal de voz $s(n)$ con una señal ruidosa $v(n)$ como:

$$y(n) = s(n) + v(n). \quad (4.8)$$

El ruido del k –ésimo coeficiente de la transformada de Fourier de tiempo corto de $y(n)$ para la ventana λ esta dada por:

$$Y_k(\lambda) = S_k(\lambda) + V_k(\lambda), \quad (4.9)$$

donde S y V son los coeficientes de la transformada de Fourier en tiempo corto para el habla limpia y con ruido respectivamente. En coordenadas polares se puede escribir como:

$$R_k(\lambda)e^{j\theta_{Y_k}(\lambda)} = A_k(\lambda)e^{j\theta_{S_k}(\lambda)} + B_k(\lambda)e^{j\theta_{V_k}(\lambda)}, \quad (4.10)$$

donde $R_k(\lambda)$, $A_k(\lambda)$ y $B_k(\lambda)$ son los espectros de magnitud de habla ruidosa, habla limpia y ruido respectivamente, $\theta_{Y_k}(\lambda)$, $\theta_{S_k}(\lambda)$ y $\theta_{V_k}(\lambda)$ son la fase de espectro de habla ruidosa, habla limpia y ruido respectivamente. Por lo tanto, la estimación del habla limpia después de la reconstrucción puede ser escrita como:

$$\widehat{S}_k(\lambda) = \widehat{A}_k(\lambda)e^{j\theta_{Y_k}(\lambda)}. \quad (4.11)$$

La magnitud del espectro de Fourier se obtiene de la siguiente ecuación:

$$|X[k]| = \sqrt{X_{re}^2 + X_{im}^2}. \quad (4.12)$$

La fase del espectro de Fourier se obtiene de la siguiente ecuación:

$$\angle X[k] = \tan^{-1} \left(\frac{X_{im}}{X_{re}} \right). \quad (4.13)$$

4.4.1 Configuración computacional para extraer los coeficientes de la transformada de Fourier en tiempo corto

Dado que el objetivo de mejorar el habla es obtener una estimación del espectro del habla, y dado que la fase no tiene un impacto perceptual significativo, este trabajo se enfoca en la reconstrucción del habla ruidosa teniendo en cuenta únicamente la fase necesaria. Del espectro de Fourier se obtiene la magnitud para mejorar la señal de voz. Se ha configurado la transformada de Fourier de tiempo corto con los siguientes parámetros:

- Una ventana de 512 muestras para calcular la transformada de Fourier en cada paso.
- Un desplazamiento de 160 muestras entre ventanas adyacentes para calcular la siguiente ventana de transformada de Fourier.
- Una longitud de ventana de análisis de 512 muestras.
- El parámetro de ventana que se utilizó para suavizar la señal fue Hamming.

4.5 Configuración del problema de adaptación de dominio

La configuración del problema de adaptación en un entorno de regresión se basa en el algoritmo que se describe en esta sección.

4.5.1 Adaptación de dominio como problema de regresión

Requerimientos de entrada: x^s (entradas del dominio fuente), y^s (etiquetas del dominio fuente), x^t (entradas del dominio objetivo), m (tamaño del lote). n_f, n_h, n_s : número de iteraciones para el entrenamiento del generador, entrenamiento del discriminador y para el entrenamiento del dominio fuente utilizando el Transporte Óptimo y n (número de iteraciones). También se requiere: θ_f (parámetros iniciales del estimador f) y θ_h (parámetros iniciales del discriminador h).

1. Para cada lote de muestras de origen (x^s, y^s) y muestras objetivo (y^t) hacer:

2. Ajustar θ_f , resolver γ de la ecuación

$$\min_{\gamma, f} \mathcal{L}_1 + \mathcal{L}_2 = \min_{\gamma, f} \frac{1}{N^s} \sum_i \|y_i^s - f(x_i^s)\|^2 + \sum_{i,j} \gamma_{ij} (\alpha \|x_i^s - x_i^t\|^2 + \beta \|y_i^s - f(x_j^t)\|^2)$$

por Transporte Óptimo.

3. Ajustar $\gamma, \theta_f \leftarrow \text{Adam}(\nabla_{\theta_f}, \mathcal{L}_2, \theta_f, \theta_h)$

4. Si n modulo $n_s == 0$ entonces

5. $\theta_f \leftarrow \text{Adam}(\nabla_{\theta_f}, \mathcal{L}_1, \theta_f, \theta_h)$

6. Si n modulo $n_f == 0$ entonces

7. $\theta_f \leftarrow \text{Adam}(\nabla_{\theta_f}, \mathcal{L}_f, \theta_f, \theta_h)$

8. Si n modulo $n_h == 0$ entonces

9. $\theta_f \leftarrow \text{Adam}(\nabla_{\theta_h}, \mathcal{L}_h, \theta_f, \theta_h)$

Se utilizó la red adversaria generativa para la tarea de reconocimiento del habla. El mapeo de características se lleva a cabo mediante un generador y un discriminador. El discriminador discrimina entre señales reales y falsas, y luego este transmite la información al generador para que este pueda aprender a producir una salida que se asemeje a la distribución realista (Phan et al. 2020). La figura 4.4 muestra la arquitectura de la red adversaria generativa utilizada para mejorar el reconocimiento del habla a través de un modelo de regresión. En la parte del generador, se emplea una red neuronal compuesta por una capa (BLSTM), seguida de dos capas completamente conectadas. Para introducir no linealidad se utilizan funciones de activación LeakyReLU y ReLU. Además, se incluye una capa de Dropout para regularizar el modelo. El objetivo del generador es tomar una entrada de tamaño 257 y generar una salida de tamaño 257, que representa la magnitud de la transformada de Fourier de tiempo corto. Por otro lado, el discriminador se implementa utilizando una arquitectura de red neuronal convolucional. Esta red consta de dos capas convolucionales seguidas de dos capas completamente conectadas. Su función es extraer características discriminativas de la magnitud del espectro de Fourier y distinguir entre muestras reales y falsas generadas por el generador.

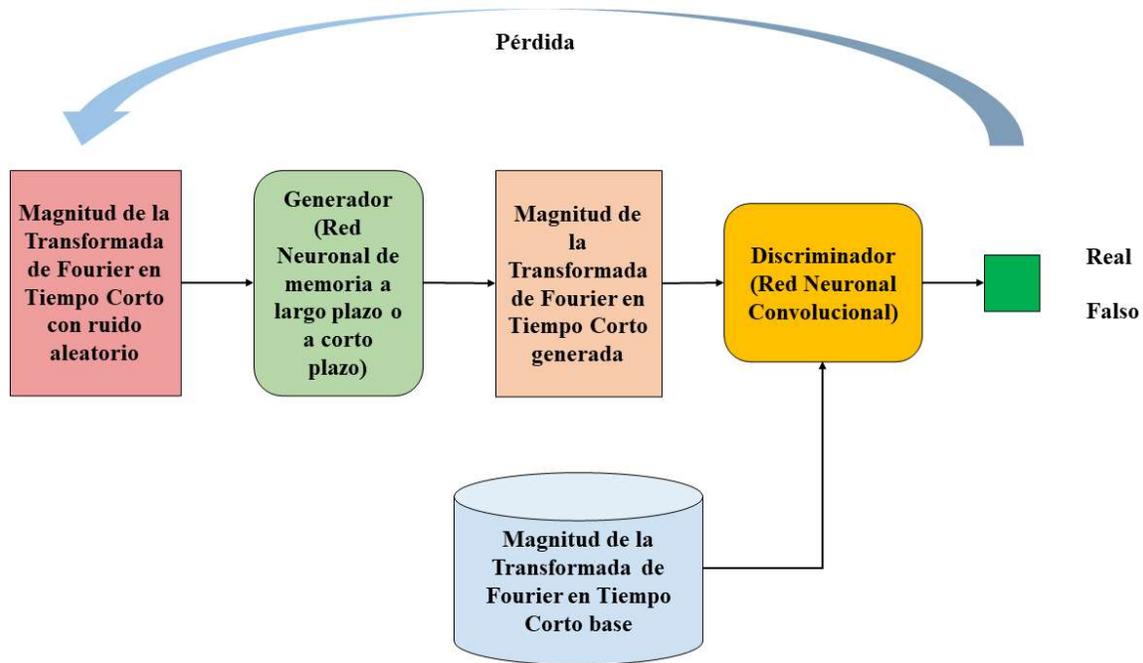


Figura 4.4 Arquitectura de red neuronal profunda con adaptación de dominio para mejora del habla ruidosa

El objetivo del modelo de regresión implica que la magnitud de la transformada de Fourier en tiempo corto sea capaz de discriminar entre aquellas con ruido y aquellas sin ruido, logrando una mayor precisión únicamente en los coeficientes limpios de la transformada de Fourier en tiempo corto.

4.5.2 Configuración computacional para la implementación de la red neuronal de mejora de la señal de voz

La tabla 4.2 muestra la arquitectura de la red neuronal utilizada para el generador en el proceso de eliminación de ruido. Durante el entrenamiento, se aplicaron los siguientes hiperparámetros: se llevaron a cabo 150 épocas de entrenamiento, se utilizaron 3 optimizadores Adam con tasas de aprendizaje correspondientes a n modulo n_f con una tasa de aprendizaje fijado en $1e-5$, n modulo n_s con una tasa de aprendizaje fijado en $1e-4$, y para ajustar γ , θ_f con una tasa de aprendizaje de $1e-5$. Además, se empleó la función de pérdida descrita en la sección 4.5.1. Para evaluar la pérdida en cada iteración, se utilizó el error cuadrático medio en cada época.

La arquitectura de la red neuronal empleada para el discriminador en el proceso de eliminación de ruido se encuentra detallada en la tabla 4.3. Durante el entrenamiento, se ajustaron los hiperparámetros de la siguiente manera: se realizaron n módulo n_h épocas de entrenamiento (se ajustaron los hiperparámetros de la red en intervalos regulares, permitiendo una adaptación continua a lo largo de las épocas de entrenamiento), se utilizó el optimizador Adam con una tasa de aprendizaje de $1e-3$, y se aplicó la función de pérdida entropía cruzada binaria.

Tabla 4.2 Arquitectura del generador para eliminación de ruido

Capa	Salida	Parámetros
Generador	(64, 257)	-
LSTM	(64, 2048)	35,692,544
Linear	(64, 1024)	2,098,176
Leaky ReLU	(64, 1024)	-
Dropout	(64, 1024)	-
Linear	(64, 257)	263,425
ReLU	(64, 257)	-
Total de parámetros	38,054,145	
Parámetros entrenables	38,054,145	

Tabla 4.3 Arquitectura del discriminador para eliminación de ruido

Capa	Salida	Parámetros
Discriminador	(1, 1)	-
Sequential 1	(1, 8, 32, 128)	-
Convolution	(1, 8, 64, 257)	208
ReLU	(1, 8, 64, 257)	-
Max pooling	(1, 8, 32, 128)	-
Sequential 2	(1, 16, 16, 64)	-
Convolution	(1, 16, 32, 128)	3,216
ReLU	(1, 16, 32, 128)	-
Max pooling	(1, 16, 16, 64)	-
Sequential 3	(1, 256)	-
Linear	(1, 256)	4,194,560
ReLU	(1, 256)	-
Sequential 4	(1, 1)	-
Linear	(1, 1)	257

(continuación)	Capa	Salida	Parámetros
	Sigmoid	(1, 1)	-
Total de parámetros		4,198,241	
Parámetros entrenables		4,198,241	

El entrenamiento de la GAN tuvo una duración total de 8 horas y 49 minutos.

4.6 Evaluación de la señal de voz

La evaluación de la inteligibilidad y calidad de la señal de voz se realizó mediante el uso de las siguientes métricas de evaluación.

4.6.1 STOI

La métrica de Inteligibilidad Objetiva de Tiempo Corto (STOI) es utilizada para evaluar las señales de habla. Esta métrica está altamente correlacionada con la inteligibilidad de las señales de habla degradadas. El valor resultante de la métrica STOI varía entre 0 y 1, donde 1 indica una inteligibilidad perfecta, mientras que 0 indica una pérdida completa de inteligibilidad (Taal et al. 2010). La tabla 4.4 muestra lo que evalúa subjetivamente STOI, así como su valor de referencia.

Tabla 4.4 Métrica STOI

¿Qué evalúa?	Puntajes de referencia
Ruido aditivo	0 - 1
Reducción de ruido	
Enmascaramiento binario	
Habla codificada	

4.6.2 PESQ

La Evaluación Perceptual de la Calidad del Habla (PESQ) es un estándar ampliamente aceptado en la industria para evaluar la calidad del audio. PESQ asigna una puntuación que varía entre -0.5 y 4.5 (Rix et al. 2001). En la tabla 4.5 se muestra lo que evalúa subjetivamente la métrica de PESQ y los puntajes de referencia.

Tabla 4.5 Métrica PESQ

¿Qué evalúa?	Puntajes de referencia
Nitidez	-0.5 – 4.5
Volumen	
Ruido de fondo	
Distorsión	
Interferencias de audio	

4.7 Configuración de la transformada Wavelet como técnica de filtrado para eliminación de ruido

La figura 4.5 muestra una técnica de filtrado conocida como umbralización Wavelet, la cual se utiliza en el proceso de eliminación de ruido de señales. En esta técnica, la señal se descompone en sub-bandas de aproximación y sub-bandas de detalles mediante la transformada Wavelet discreta. Por lo general, los componentes de ruido se concentran en las sub-bandas de alta frecuencia, mientras que la información principal se encuentra en la sub-banda de baja frecuencia (Thu et al. 2019).

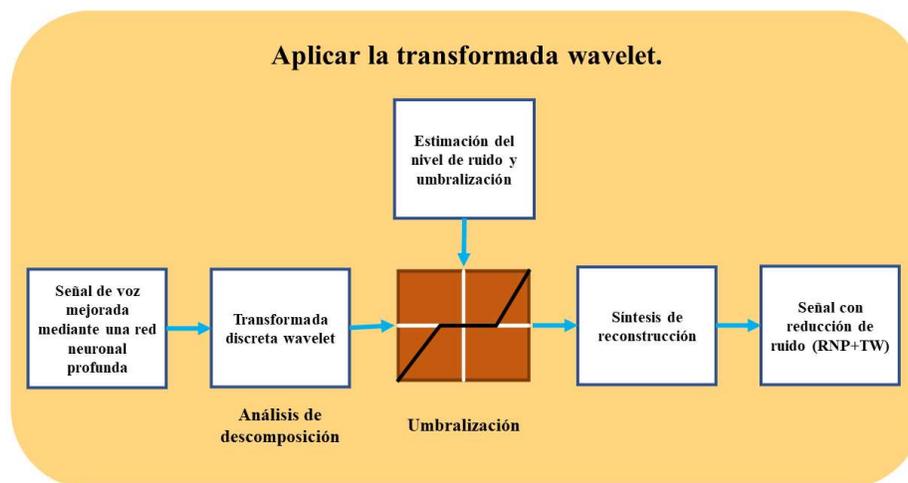


Figura 4.5 Técnica de filtrado basado en transformada Wavelet

La elección de una función Wavelet adecuada es crucial en el proceso de eliminación de ruido, ya que el uso de funciones inapropiadas puede hacer que la transformada sea compleja e inmanejable. En este caso particular, se probó la configuración con una función Wavelet del tipo Daubechies. Además, se utilizó un número de niveles de descomposición Wavelet igual a 9. El método seleccionado para realizar la eliminación de ruido fue el método VisuShrink; este método permite eliminar los coeficientes que se consideran ruido.

4.8 Modelos del reconocimiento del habla

Se muestra en la figura 4.6 el diagrama de bloques utilizado para evaluar el habla base, el habla ruidosa y el habla después de la eliminación de ruido. Este diagrama consta de dos bloques principales. El primer bloque corresponde al modelo de identificación de locutor, mientras que el segundo bloque se encarga del RAV.

4.8.1 Configuración del modelo de identificación de locutor

En este bloque se separan los audios en conjuntos de entrenamiento y prueba (ver figura 4.6). Se seleccionaron 10 locutores, cada uno con 100 audios. A continuación, se extraen las características del habla utilizando coeficientes cepstrales en frecuencia de Mel. Es importante

destacar que para este modelo se utilizan únicamente 5 segundos de cada audio, lo que puede implicar la adición o eliminación de fragmentos de audio según corresponda. Las matrices resultantes de características se utilizan para alimentar el modelo de aprendizaje profundo diseñado para el reconocimiento de locutor. Una vez que el modelo ha pasado por cada época de entrenamiento, se evalúa su rendimiento en la identificación de locutor utilizando métricas como exactitud, recall, F1-score y especificidad.

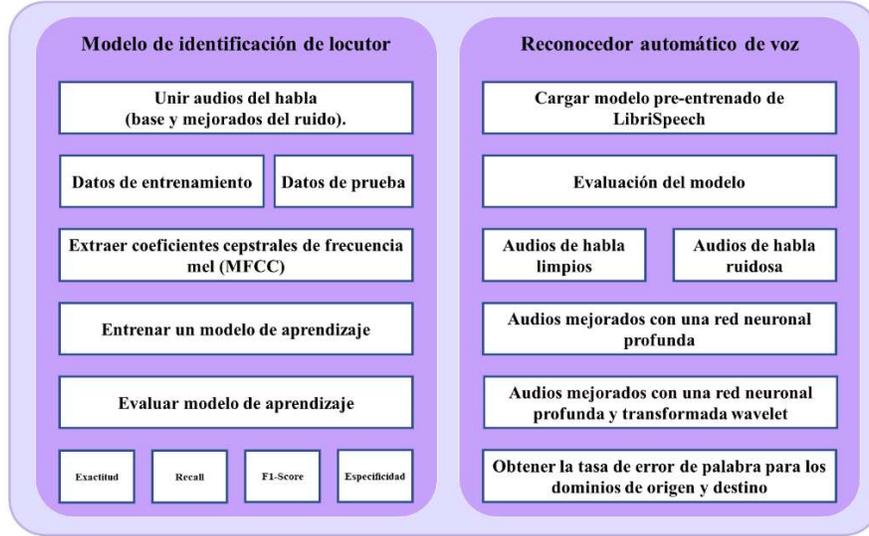


Figura 4.6 Modelos del reconocimiento del habla

4.8.1.1 Adaptación de dominio como problema de clasificación

La configuración del problema de adaptación bajo un entorno de clasificación se basa en el algoritmo que se describe en esta sección. Requerimientos de entrada: x^s (entradas del dominio fuente), y^s (etiquetas del dominio fuente), x^t (entradas del dominio objetivo), m (tamaño del lote); n_c, n_f, n_h, n_s , que es el número de iteraciones del Transporte Óptimo para el entrenamiento del clasificador, entrenamiento del generador, entrenamiento del discriminador y para el entramiento del dominio fuente, así como n (número de iteraciones). También se requiere: θ_c (parámetros iniciales del clasificador c), θ_f (parámetros iniciales del estimador f) y θ_h , (parámetros iniciales del discriminador h).

1. Para cada lote de muestras de origen (x^s, y^s) y muestras objetivo (y^t) hacer:
2. Ajustar θ_c y θ_f , resolver por separado γ de la ecuación
$$\min_{\gamma, f} \mathcal{L}_1 + \mathcal{L}_2 = \min_{\gamma, f} \frac{1}{N^s} \sum_i \|y_i^s - f(x_i^s)\|^2 + \sum_{i,j} \gamma_{ij} (\alpha \|x_i^s - x_j^t\|^2 + \beta \|y_i^s - f(x_j^t)\|^2)$$
por Transporte Óptimo.
3. Ajustar $\gamma, \theta_c \leftarrow \text{Adam}(\nabla_{\theta_c}, \mathcal{L}_2, \theta_c, \theta_h)$
4. Ajustar $\gamma, \theta_f \leftarrow \text{Adam}(\nabla_{\theta_f}, \mathcal{L}_2, \theta_f, \theta_h)$
5. Si n modulo $n_s == 0$ entonces

6. $\theta_f \leftarrow Adam(\nabla_{\theta_f}, \mathcal{L}_1, \theta_f, \theta_h)$
7. Si n modulo $n_f == 0$ entonces
8. $\theta_f \leftarrow Adam(\nabla_{\theta_f}, \mathcal{L}_f, \theta_f, \theta_h)$
9. Si n modulo $n_h == 0$ entonces
10. $\theta_f \leftarrow Adam(\nabla_{\theta_h}, \mathcal{L}_h, \theta_f, \theta_h)$

La figura 4.7 muestra la arquitectura de la red adversaria generativa utilizada para la identificación de locutor como un modelo de clasificación. El generador de la red adversaria generativa se compone de una red neuronal de una capa BLSTM, seguida de dos capas completamente conectadas. Se emplean funciones de activación LeakyReLU y ReLU, junto con una capa de Dropout para la regularización del modelo. El propósito de este generador es tomar una entrada de tamaño 40 y producir una salida de tamaño 40, correspondiente a la cantidad de MFCCs extraídos. Por otro lado, el discriminador se implementa utilizando una arquitectura de red neuronal convolucional. Esta red consta de dos capas convolucionales, seguidas de dos capas completamente conectadas. Su función principal es extraer características discriminativas de los MFCCs y realizar la discriminación entre muestras reales y falsas generadas por el generador. Además, para la clasificación del locutor se utiliza una red neuronal convolucional denominada VGG, que cuenta con 13 capas convolucionales y 3 capas completamente conectadas.

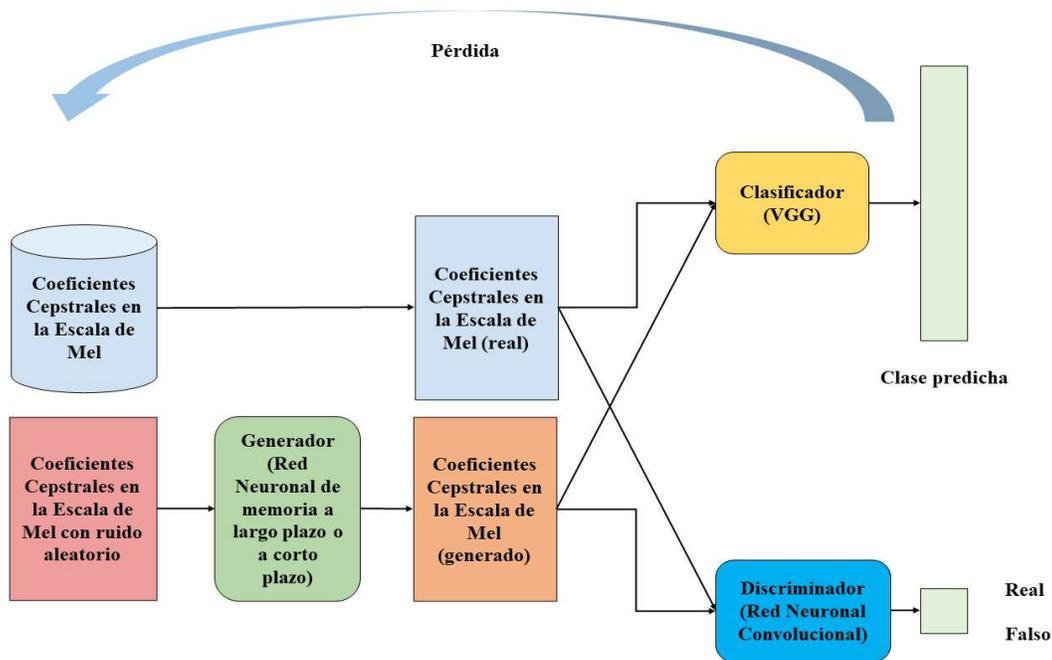


Figura 4.7 Arquitectura de red neuronal profunda para clasificación de locutor

El objetivo del clasificador implica que los MFCC sean capaces de discriminar entre aquellos con ruido y aquellos sin ruido, obteniendo una mayor precisión únicamente en los MFCC limpios.

4.8.1.2 Configuración computacional para extraer los coeficientes cepstrales en frecuencia de Mel

Para identificación de locutor se utilizó el proceso del método de extracción de características basado en Coeficientes Cepstrales en Frecuencia de Mel (MFCC) (Daniel Jurafsky, 2009). Se extrajeron 40 MFCCs y 40 parámetros específicos para el mel-espectrograma. Estos parámetros incluyen un tamaño de transformada de Fourier de tiempo corto de 256, una longitud de salto entre ventanas adyacentes calculada del 25% del tamaño de la transformada, un número de bandas de Mel calculado como la mitad del tamaño de la transformada más uno. Es importante señalar que se acotaron los audios en un tiempo de 3 segundos para extraer los MFCCs.

4.8.1.3 Configuración computacional para la implementación de la red neuronal de identificación de locutor

En la tabla 4.6 se detalla la configuración de los parámetros de la red neuronal correspondiente al generador. Esta red neuronal fue entrenada a lo largo de 150 épocas además de ajustar los parámetros cada n módulo n_s , utilizando el optimizador Adam con una tasa de aprendizaje de $1e-5$. La función de pérdida empleada sigue el algoritmo propuesto en la sección 4.8.1.1.

Tabla 4.6 Arquitectura del generador

Capa	Salida	Parámetros
Generador	(512, 40)	-
LSTM	(512, 1024)	8,568,832
Linear	(512, 512)	524,800
Leaky ReLU	(512, 512)	-
Dropout	(512, 512)	-
Linear	(512, 40)	20,520
ReLU	(512, 40)	-
Total de parámetros	9,114,152	
Parámetros entrenables	9,114,152	

La tabla 4.7 presenta la arquitectura de la red neuronal correspondiente al discriminador, donde se detallan los parámetros específicos de entrada y salida. Esta red neuronal fue entrenada en intervalos de n módulo n_h épocas, utilizando el método Adam con una tasa de aprendizaje de $1e-3$. Para esta red neuronal, se empleó la función de pérdida de entropía cruzada binaria.

Tabla 4.7 Arquitectura del discriminador

Capa	Salida	Parámetros
Discriminador	(1, 1)	-
Sequential 1	(1, 8, 20, 256)	-
Convolution	(1, 8, 40, 512)	208
ReLU	(1, 8, 40, 512)	-
Max pooling	(1, 8, 20, 256)	-
Sequential 2	(1, 16, 10, 128)	-
Convolution	(1, 16, 20, 256)	3,216
ReLU	(1, 16, 20, 256)	-
Max pooling	(1, 16, 10, 128)	-
Sequential 3	(1, 256)	-
Linear	(1, 256)	5,243,136
ReLU	(1, 256)	-
Sequential 4	(1, 1)	-
Linear	(1, 1)	257
Sigmoid	(1, 1)	-
Total de parámetros	5,246,817	
Parámetros entrenables	5,246,817	

La tabla 4.8 muestra los parámetros de la red neuronal VGG-16, donde se detallan tanto los parámetros de entrada como los de salida correspondientes. Durante el proceso de entrenamiento de esta red neuronal, se llevaron a cabo 150 épocas. Se utilizó la función de pérdida descrita en el algoritmo de la sección 4.8.1.1, y se ajustaron los parámetros de la red neuronal mediante el método Adam, con una tasa de aprendizaje de $1e-5$. La pérdida fue medida utilizando la entropía cruzada, y se calculó la exactitud para evaluar su desempeño en la clasificación en cada época.

Tabla 4.8 Arquitectura de red neuronal para clasificador de locutores

Capa	Salida	Parámetros
VGG16	(1, 10)	-
Sequential 1	(1, 64, 40, 512)	-
Convolution	(1, 64, 40, 512)	-
Batch normalization	(1, 64, 40, 512)	640
ReLU	(1, 64, 40, 512)	128
Sequential 2	(1, 64, 20, 256)	-
Convolution	(1, 64, 40, 512)	-
Batch normalization	(1, 64, 40, 512)	36,928
ReLU	(1, 64, 40, 512)	128
Max pooling	(1, 64, 20, 256)	-
Sequential 3	(1, 128, 20, 256)	-
Convolution	(1, 128, 20, 256)	-
Batch normalization	(1, 128, 20, 256)	73,856
ReLU	(1, 128, 20, 256)	256
Sequential 4	(1, 128, 10, 128)	-
Convolution	(1, 128, 20, 256)	-
Batch normalization	(1, 128, 20, 256]	147,584
ReLU	(1, 128, 20, 256]	256
Max pooling	(1, 128, 10, 128)	-
Sequential 5	(1, 256, 10, 128)	-
Convolution	(1, 256, 10, 128)	-
Batch normalization	(1, 256, 10, 128)	295,168
ReLU	(1, 256, 10, 128)	512
Sequential 6	(1, 256, 10, 128)	-
Convolution	(1, 256, 10, 128)	-
Batch normalization	(1, 256, 10, 128)	590,080

(continuación)	Capa	Salida	Parámetros
	ReLU	(1, 256, 10, 128)	512
Sequential 7		(1, 256, 5, 64)	-
	Convolution	(1, 256, 10, 128)	-
	Batch normalization	(1, 256, 10, 128)	590,080
	ReLU	(1, 256, 10, 128)	512
	Max pooling	(1, 256, 5, 64)	-
Sequential 8		(1, 512, 5, 64)	-
	Convolution	(1, 512, 5, 64)	1,180,160
	Batch normalization	(1, 512, 5, 64)	1,024
	ReLU	(1, 512, 5, 64)	-
Sequential 9		(1, 512, 5, 64)	-
	Convolution	(1, 512, 5, 64)	2,359,808
	Batch normalization	(1, 512, 5, 64)	1,024
	ReLU	(1, 512, 5, 64)	-
Sequential 10		(1, 512, 2, 32)	-
	Convolution	(1, 512, 5, 64)	2,359,808
	Batch normalization	(1, 512, 5, 64)	1,024
	ReLU	(1, 512, 5, 64)	-
	Max pooling	(1, 512, 2, 32)	-
Sequential 11		(1, 512, 2, 32)	-
	Convolution	(1, 512, 2, 32)	2,359,808
	Batch normalization	(1, 512, 2, 32)	1,024
	ReLU	(1, 512, 2, 32)	-
Sequential 12		(1, 512, 2, 32)	-
	Convolution	(1, 512, 2, 32)	2,359,808
	Batch normalization	(1, 512, 2, 32)	1,024
	ReLU	(1, 512, 2, 32)	-

(continuación)	Capa	Salida	Parámetros
Sequential 13		(1, 512, 1, 16)	-
	Convolution	(1, 512, 2, 32)	2,359,808
	Batch normalization	(1, 512, 2, 32)	1,024
	ReLU	(1, 512, 2, 32)	-
	Max pooling	(1, 512, 1, 16)	-
Sequential 15		(1, 4096)	-
	Dropout	(1, 8192)	-
	Linear	(1, 4096)	33,558,528
	ReLU	(1, 4096)	-
Sequential 15		(1, 4096)	-
	Dropout	(1, 4096)	-
	Linear	(1, 4096)	16,781,312
	ReLU	(1, 4096)	-
Sequential 16		(1, 10)	-
	Linear	(1, 10)	40,970
Total de parámetros		65,102,794	
Parámetros entrenables		65,102,794	

El entrenamiento de la red neuronal para la identificación de locutor tuvo una duración promedio de 5 horas y 58 minutos.

4.8.1.4 Métricas de evaluación del modelo de clasificación

En este apartado se definen las métricas de evaluación para el modelo de clasificación (Geron, 2017).

- Exactitud es la fracción de predicciones que fueron correctas por nuestro modelo.

$$\text{Exactitud} = \frac{\text{Numero de predicciones correctas}}{\text{Numero total de predicciones}} \quad (4.14)$$

- Recall se define como la proporción de negativos reales que se predijo que serían negativos. Esto implica que habrá otra proporción de verdaderos negativos que fueron pronosticados como positivos y podrían llamarse falsos positivos.

$$\text{Recall} = \frac{\text{Verdadero Positivo}}{\text{Verdadero Positivo} + \text{Falso Negativo}}. \quad (4.15)$$

- Especificidad es una medida de proporción de casos positivos reales que se pronosticaron como positivos. Esto implica que habrá otra proporción de casos positivos reales que se predecirían incorrectamente como negativos.

$$\text{Especificidad} = \frac{\text{Verdadero Negativo}}{\text{Verdadero Negativo} + \text{Falso Positivo}}. \quad (4.16)$$

- F1-Score es una métrica que mide la precisión de un modelo. Combina las puntuaciones de precisión y recall de un modelo. La métrica de precisión calcula cuántas veces un modelo hizo una predicción correcta en todo el conjunto de datos.

$$\text{F1 - Score} = \frac{2 \cdot \text{Exactitud} \cdot \text{Recall}}{\text{Exactitud} + \text{Recall}}. \quad (4.17)$$

4.8.2 Configuración del modelo reconocedor automático de voz

Para el RAV se cuenta con un reconocedor que ha sido entrenado previamente por Watanabe (2021) utilizando el corpus de LibriSpeech. Se generó un subconjunto de datos de LibriSpeech compuesto por 10 locutores. De estos 10 locutores, se seleccionaron aleatoriamente 100 audios para probar el modelo de eliminación de ruido. Estos conjuntos de datos se dividen en cuatro grupos: el primero corresponde al habla base, el segundo al habla ruidosa, el tercero al habla después de aplicar únicamente la red neuronal para la eliminación de ruido, y el último al habla procesada con la red neuronal y la transformada Wavelet. Estos conjuntos de datos se introducen en el modelo de reconocimiento automático de voz y se obtiene el rendimiento mediante la tasa de error de palabra.

4.8.2.1 Evaluación del modelo reconocedor automático de voz

La tasa de error de palabra (WER) es la métrica de evaluación estándar para los sistemas de reconocimiento de voz. Esta métrica se basa en cuánto difiere la cadena de palabras devuelta por el reconocedor de una transcripción correcta o de referencia (Daniel Jurafsky, 2009).

$$\text{WER} = 100 \cdot \frac{\text{Inserciones} + \text{Eliminaciones} + \text{Sustituciones}}{\text{Total de palabras}}. \quad (4.18)$$

donde dada una referencia de palabras y una hipótesis de palabras del reconocedor:

Inserciones: Número de palabras insertadas en la secuencia de palabras de referencia.

Eliminaciones: Número de palabras eliminadas en la secuencia de palabras de referencia.

Sustituciones: Número de palabras sustituidas en la secuencia de palabras de referencia.

4.9 Hardware y software utilizado

4.9.1 Recursos de hardware

La tabla 4.9 detalla el equipo utilizado para la experimentación.

Tabla 4.9 Especificaciones del equipo

Modelo del equipo: Laptop Alienware M15 R6	
Componente	Capacidad
Sistema operativo	Ubuntu 22.04.2 LTS
RAM	16 GB DDR4 3200 Hz
Procesador	11th Gen Intel (R) Core (TM) i7-11800H 2.30GHz
Tarjeta grafica	NVIDIA GeForce RTX 3060

4.9.2 Recursos de software

En esta sección se detallan los recursos de software utilizados para llevar a cabo la implementación de la experimentación. La tabla 4.10 presenta una lista de las bibliotecas empleadas para cada componente utilizado en la investigación. Es importante destacar que toda la experimentación se realizó dentro del entorno de programación Python.

Tabla 4.10 Software empleado para la experimentación

Componente	Librería
Red neuronal para regresión	Pytorch
Transformada Wavelet	scikit-image
Red neuronal para clasificación	Pytorch
Reconocedor automático de voz	EspNet
Transformada de Fourier de Tiempo Corto	Librosa
Coeeficientes Cepstrales en las Frecuencias de Mel	Librosa
Métricas de evaluación de los modelos	Torchmetrics
Plan de Transporte Óptimo	POT

Además, se emplearon bibliotecas como matplotlib, numpy, pandas, pickle, shutil, scipy, math, gc, os y random para propósitos generales dentro del desarrollo de la implementación de la experimentación.

Capítulo 5 Resultados y limitaciones

En este capítulo se presentan los resultados obtenidos a partir de la propuesta de investigación. Se comienza con el análisis del comportamiento de la función de pérdida de la red neuronal propuesta para la eliminación de ruido. A continuación, se exponen los resultados particulares de la eliminación de ruido en la señal de voz, tanto en su forma de onda como en su espectrograma. Posteriormente, se lleva a cabo un análisis general de la eliminación de ruido, abordando cada tipo de ruido y en sus diferentes niveles de relación señal-ruido. Se evalúan todas las métricas correspondientes, centrándonos en analizar dichas métricas cuando la señal contiene ruido, cuando se mejora con la red neuronal y cuando se aplica tanto la red neuronal como la técnica de filtrado basada en la transformada Wavelet. A continuación, se presenta cómo los resultados de las métricas de intangibilidad y calidad se reflejan en un sistema de reconocimiento de voz automático, así como en la identificación de locutor. De esta manera, se muestra el impacto de la eliminación de ruido en el rendimiento de estos sistemas de reconocimiento del habla.

5.1 Comportamiento de la función de pérdida en los dominios fuente y objetivo

La figura 5.1 representa el comportamiento de la función de pérdida durante el entrenamiento de la red neuronal en el dominio fuente, donde se logró obtener un error cuadrático medio (ECM) de 0.00621. Por otro lado, la figura 5.2 muestra el comportamiento de la función de pérdida en el dominio destino, con un ECM de 0.00893. En ambas figuras, se puede observar cómo la función de pérdida de la red neuronal evoluciona a lo largo de las épocas tanto en los datos de entrenamiento como en los datos de prueba. Estas figuras reflejan el rendimiento y la capacidad de la red neuronal para mejorar el habla ruidosa. A medida que el entrenamiento progresa, la función de pérdida disminuye, lo que indica que la red neuronal está aprendiendo y ajustando sus parámetros para minimizar la diferencia entre la salida deseada y la salida real. La comparación entre la figura 5.1 y la figura 5.2 muestra cómo la red neuronal se comporta en el dominio fuente u origen y en el dominio destino, respectivamente.

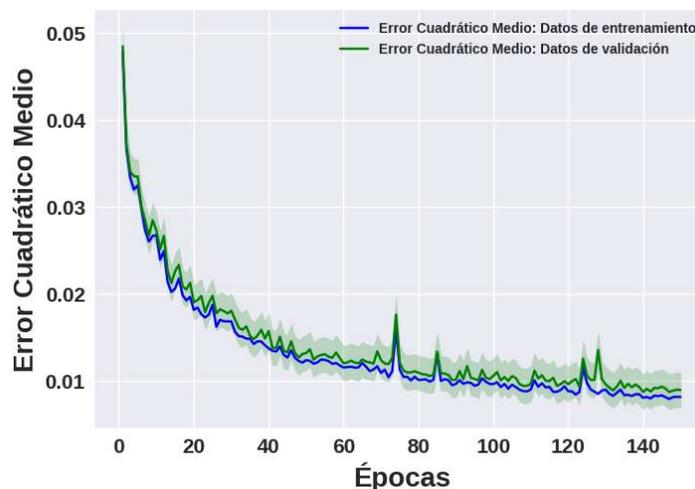


Figura 5.1 Comportamiento de función de pérdida ECM para mejora del habla en el dominio origen

En la figura 5.2 se observa un grado moderado de degradación en el ruido que se intenta eliminar en el dominio objetivo. A pesar de esta degradación moderada, la red neuronal muestra su capacidad para eliminar el ruido en ambos dominios. A pesar de que la señal de voz con ruido presenta cierta pérdida de calidad en el dominio objetivo, la red neuronal es capaz de realizar una reducción efectiva del ruido, mejorando la calidad general de la señal. Esto demuestra la capacidad de la red neuronal para aprender y adaptarse a las características del ruido presente en el dominio objetivo, lo que a su vez contribuye a la eliminación del ruido no deseado.

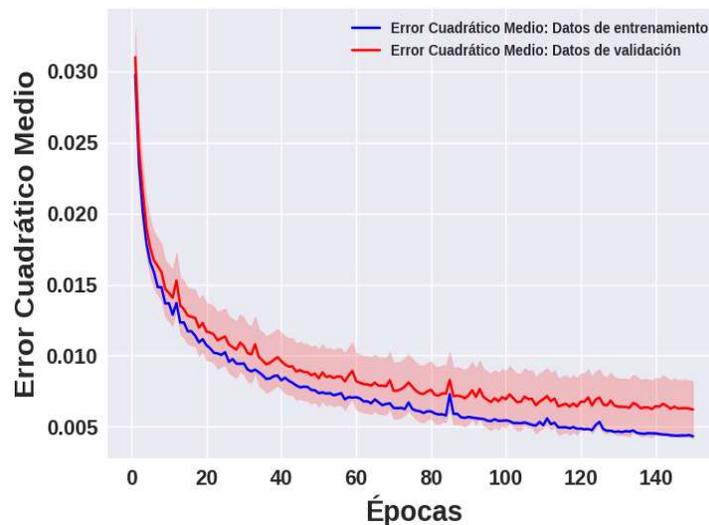


Figura 5.2 Comportamiento de función de pérdida ECM para mejora del habla en el dominio destino

En la figura 5.3 se presentan cuatro casos de una señal de audio en el dominio del tiempo. En la figura 5.3 (a) se muestra la señal de voz de un locutor en su forma base, sin ninguna alteración. En la figura 5.3 (b) se muestra la misma señal con ruido añadido. En este caso, el ruido añadido es de naturaleza no estacionaria y se le asignó un nivel de relación señal-ruido (SNR) de -3dB. Como se puede observar en la figura, la presencia del ruido hace que la señal de voz sea prácticamente indistinguible. Cuando una señal de voz con este tipo de ruido se somete a un sistema de reconocimiento de voz automático, el rendimiento del sistema tiende a degradarse en términos de inteligibilidad y comprensión, lo que resulta en un rendimiento inferior. Finalmente, en la figura 5.3 (c) y figura 5.3 (d) se muestra la señal de voz después de aplicar dos técnicas diferentes para eliminar el ruido: una red neuronal propuesta y una técnica de filtrado basada en la transformada Wavelet. En ambos casos, se puede observar cómo los componentes de la señal a lo largo del tiempo se vuelven más definidos y se recupera la calidad original de la señal de voz. Esta mejora es crucial para que los sistemas de RAV tengan un rendimiento óptimo, ya que se facilita la identificación y el análisis de las características acústicas clave en la señal.

En la figura 5.4 se muestran los espectrogramas del espectro de Fourier. El espectrograma representa la variación de la energía en función del tiempo y la frecuencia. El espectrograma, además de mostrar dicha información, también es una herramienta para visualizar propiedades de una señal de audio, como la presencia de tonos, armónicos y cambios en el contenido de frecuencia a lo largo de tiempo. El espectrograma de una señal de voz base tomada del corpus de LibreSpeech se representa en la figura 5.4 (a). Este espectrograma muestra cómo varía la cantidad de energía

en cada componente de frecuencia a lo largo del tiempo. Los colores brillantes en el espectrograma indican una mayor concentración de energía, mientras que los colores claros indican una menor energía. En el caso del reconocimiento del habla, es crucial tener un espectrograma bien definido para extraer características acústicas relevantes de la señal de voz. Esto implica eliminar el ruido contenido en la señal de voz. En la figura 5.4 (b) se muestra el espectrograma de la señal de voz base cuando se le ha añadido algún tipo de ruido. Esto resulta en una representación con poca definición de los componentes de frecuencia a lo largo del tiempo. El ruido agregado ocupa todas las bandas de frecuencia en diferentes momentos del tiempo, lo que reduce la claridad y la percepción del habla en el espectrograma. La presencia de ruido tiene un impacto negativo en la calidad y la inteligibilidad de la señal de voz, lo que dificulta el reconocimiento preciso del habla.

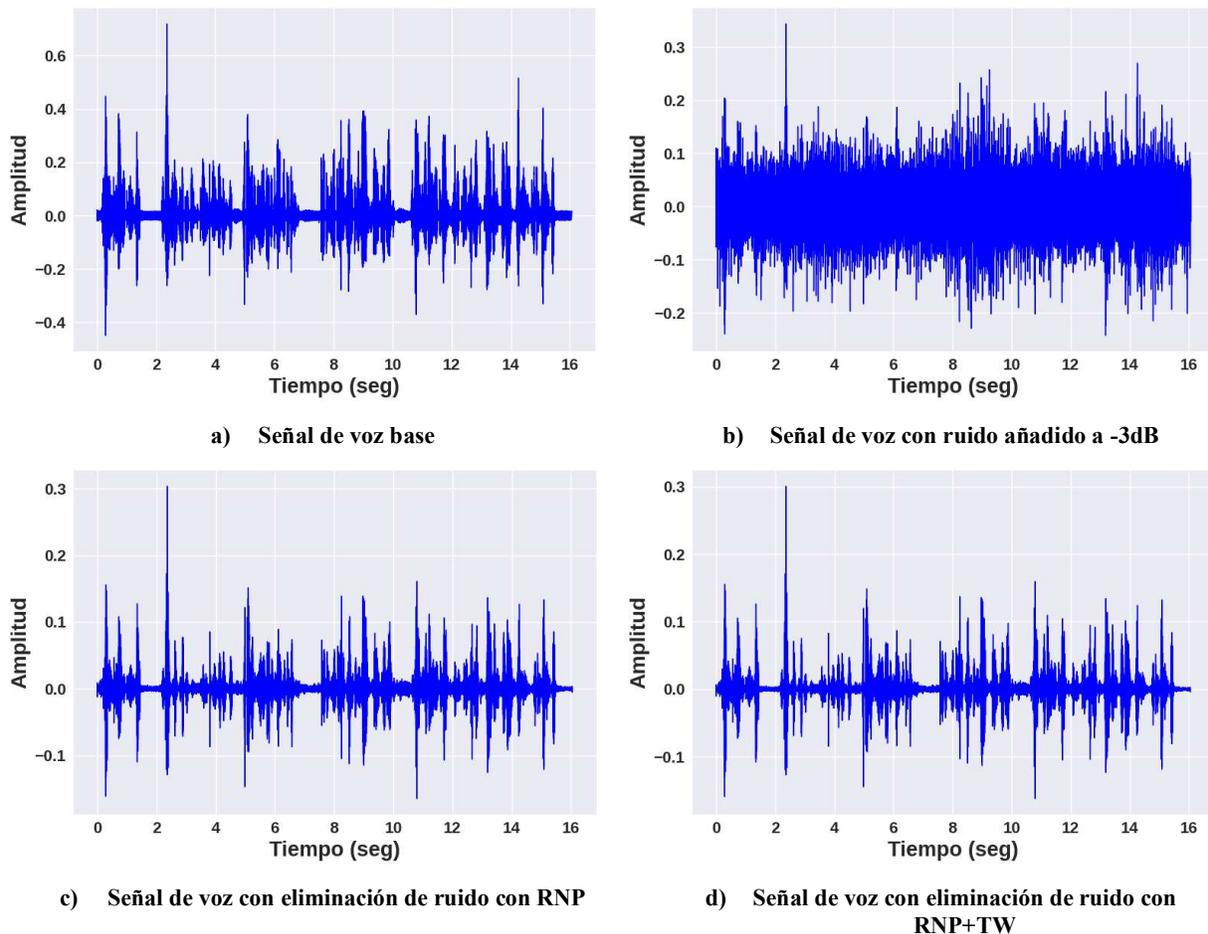


Figura 5.3 Formas de onda de la señal de voz en el dominio del tiempo para un enunciado pronunciado por un locutor del conjunto de datos LibriSpeech en su versión base. Esta señal está contaminada con el ruido de helicóptero proporcionado en un SNR nivel -3 dB y sus versiones mejoradas

Por tanto, es crucial aplicar técnicas de eliminación de ruido en el procesamiento de la señal de voz. Estas técnicas ayudan a mejorar la calidad del espectrograma, lo que a su vez tiene un efecto directo en la precisión del reconocimiento del habla. Al eliminar o reducir el ruido no deseado, se logra obtener un espectrograma más nítido y definido, lo que facilita la extracción de las características acústicas fundamentales para el reconocimiento adecuado del habla. Esto

permite una mejor identificación de los componentes de frecuencia relevantes y una representación más clara de su evolución en el tiempo. Finalmente, en la figura 5.4 (c) y figura 5.4 (d) se muestran los espectrogramas con el ruido de la señal de voz eliminado con la red neuronal profunda y después de aplicar la transformada Wavelet. Podemos notar en la figura 5.4 (c) cómo los componentes en frecuencia están más definidos a lo largo del tiempo. Es importante señalar que durante el proceso de eliminación de ruido se puede suprimir energía de la señal, tal es el caso que se muestra marcado en la figura 5.4 (c) con los recuadros verdes, el cual respecto a la figura 5.4 (a) existe una supresión de energía. Por otro lado, al aplicar la transformada Wavelet se trata de restaurar la energía suprimida, esto se ve reflejado en la métrica de PESQ, el cual ayuda a mejorar gradualmente la calidad de la señal de voz. Al reducir o eliminar el ruido no deseado, se puede obtener un espectrograma más limpio y definido, lo que facilita la extracción de características acústicas relevantes para el reconocimiento del habla.

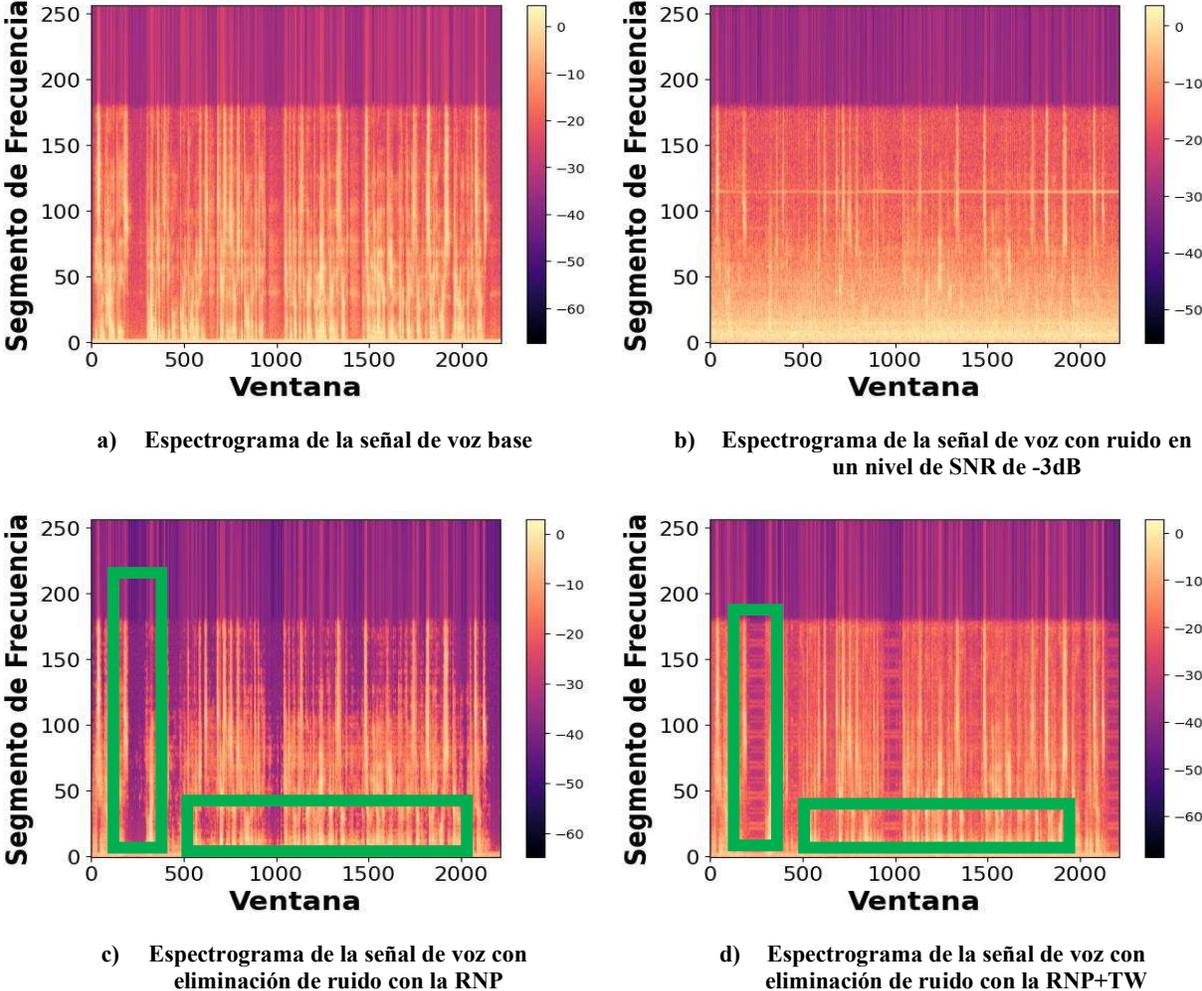


Figura 5.4 Comportamiento del espectro de la señal de voz para un enunciado pronunciado por un locutor de LibreSpeech en su versión base. Esta señal está contaminada con el ruido de helicóptero proporcionado en un SNR nivel -3 dB y sus versiones mejoradas

5.2 Resultados de eliminación de ruido en la señal de voz del dominio fuente

En esta sección se analizan los resultados obtenidos para la eliminación de ruido de tipo estacionario en el dominio fuente. Se realizaron mediciones STOI y PESQ utilizando tres enfoques diferentes: habla con ruido, habla con eliminación de ruido mediante una red neuronal profunda, y habla con eliminación de ruido utilizando tanto una red neuronal profunda como la transformada Wavelet. Estos resultados obtenidos van encaminados a los obtenidos durante el entrenamiento del dominio destino de la figura 5.1, el cual nos indica que durante la fase de entrenamiento de la red neuronal fue capaz de mitigar los distintos tipos de ruido propuestos para el dominio fuente.

5.2.1 Resultados para tipo de ruido estacionario – ruido rosa

En la figura 5.5 se presentan las evaluaciones de las métricas STOI y PESQ para el tratamiento del ruido rosa. En la figura 5.5 (a) se muestra el comportamiento del STOI en distintos niveles de relación señal-ruido (SNR). Por otro lado, en la figura 5.5 (b) se puede apreciar el comportamiento del PESQ en diferentes niveles de SNR.

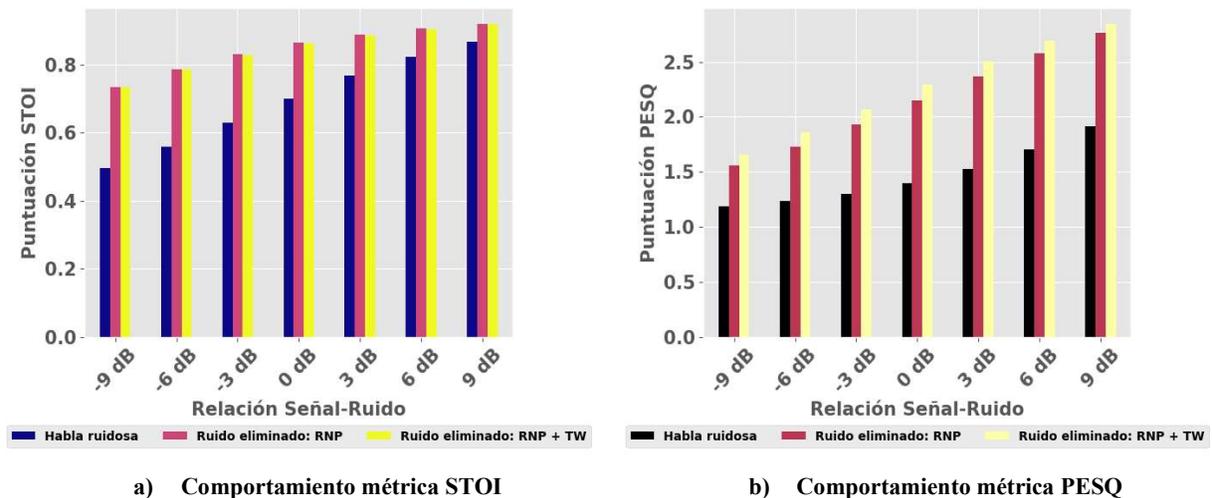


Figura 5.5 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: rosa

Según los datos presentados en la tabla 5.1, cuando una señal de voz contiene ruido rosa, la métrica STOI arroja un promedio de 0.6914. Sin embargo, al someter la señal de voz a la red neuronal profunda, esta métrica mejora significativamente a 0.8473. Por otro lado, al aplicar la transformada Wavelet a la señal procesada por la red neuronal, se observa una degradación mínima de 0.0023 en esta métrica, lo cual representa una pérdida poco relevante. En cuanto a la métrica de PESQ, cuando una señal contiene ruido, el puntaje promedio es de 1.4660. Sin embargo, al someter la señal a la red neuronal profunda, este puntaje mejora considerablemente a 2.1533. Finalmente, al aplicar la transformada a la señal obtenida de la red neuronal, se logra una mejora adicional en la métrica de 2.2734.

Tabla 5.1 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: rosa

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.4963	0.7345	0.7336	1.1864	1.5604	1.6550
-6 dB	0.5587	0.7870	0.7859	1.2311	1.7269	1.8555
-3 dB	0.6291	0.8315	0.8284	1.2991	1.9308	2.0651
0 dB	0.6991	0.8632	0.8605	1.3961	2.1464	2.2918
3 dB	0.7666	0.8882	0.8854	1.5291	2.3686	2.5066
6 dB	0.8227	0.9063	0.9034	1.7026	2.5773	2.6910
9 dB	0.8673	0.9206	0.9183	1.9181	2.7632	2.8490
Promedio	0.6914	0.8473	0.8450	1.4660	2.1533	2.2734

5.2.2 Resultados para tipo de ruido estacionario – gotas de agua

En la figura 5.6 se muestran las evaluaciones de las métricas STOI y PESQ para el tratamiento del ruido de gotas de agua. En la figura 5.6 (a) se presenta el comportamiento del STOI en distintos niveles de relación señal-ruido (SNR) específicamente para este tipo de ruido. Por otro lado, en la figura 5.6 (b) se puede observar el comportamiento del PESQ en diferentes niveles de SNR al enfrentar la señal de voz con el ruido de gotas de agua.

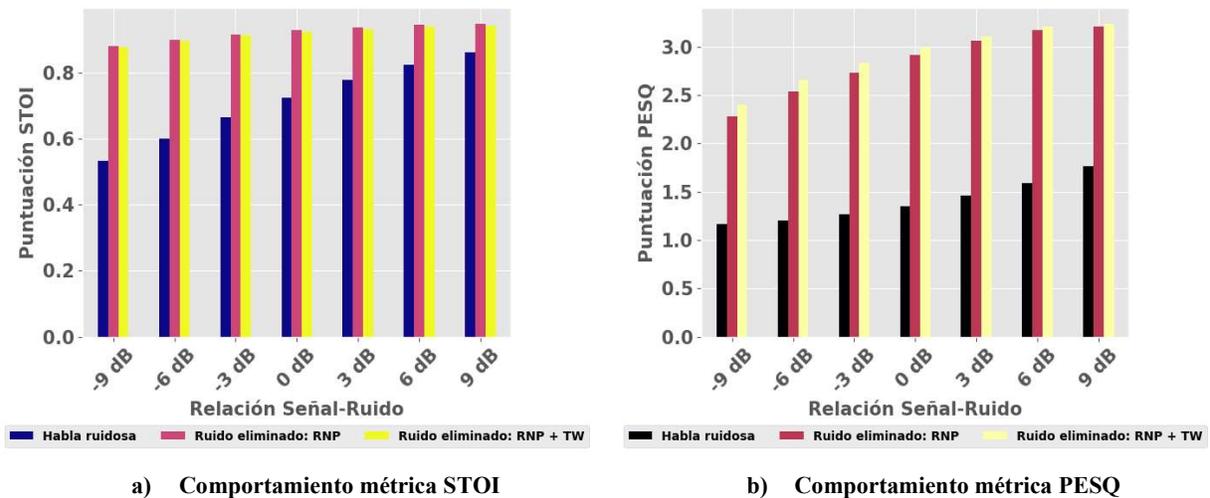


Figura 5.6 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: gotas de agua

De acuerdo con los datos presentados en la tabla 5.2, se observa que cuando una señal de voz contiene ruido de gotas de agua, la métrica STOI tiene un promedio de 0.7118. No obstante, al aplicar la señal de voz a la red neuronal profunda, se logra una mejora significativa en esta métrica, alcanzando un valor promedio de 0.9218. Al aplicar la transformada Wavelet a la señal procesada por la red neuronal, se aprecia una ligera degradación de la métrica en un valor mínimo de 0.0044, lo cual se considera una pérdida poco relevante. En cuanto a la métrica de PESQ, cuando una señal

contiene ruido de gotas de agua, el puntaje promedio es de 1.4022. Sin embargo, al someter la señal a la red neuronal profunda, se produce una mejora significativa en esta métrica, con un puntaje promedio de 2.8440. Finalmente, al aplicar la transformada a la señal obtenida de la red neuronal, se logra una mejora adicional en la métrica de 2.9182.

Tabla 5.2 Puntajes STOI y PESQ para corpus LibreSpeech - Tipo de ruido: Gotas de agua

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.5320	0.8801	0.8774	1.1661	2.2796	2.3983
-6 dB	0.5999	0.9003	0.8973	1.2060	2.5343	2.6543
-3 dB	0.6652	0.9158	0.9121	1.2693	2.7339	2.8325
0 dB	0.7242	0.9275	0.9232	1.3500	2.9136	2.9885
3 dB	0.7781	0.9365	0.9316	1.4634	3.0636	3.1068
6 dB	0.8227	0.9440	0.9383	1.5934	3.1741	3.2059
9 dB	0.8609	0.9489	0.9421	1.7677	3.2094	3.2417
Promedio	0.7118	0.9218	0.9174	1.4022	2.8440	2.9182

5.2.3 Resultados para tipo de ruido estacionario – carro

En la figura 5.7 se presentan los resultados obtenidos al aplicar la eliminación de ruido de fondo en la señal de voz. El análisis se divide en dos sub-figuras, donde se examina el comportamiento de dos métricas distintas en función de los diferentes niveles de relación señal-ruido (SNR). En la figura 5.7 (a) se muestra el comportamiento de la métrica STOI para la señal de voz en relación con los distintos niveles de SNR. Esta métrica permite evaluar la mejora de la inteligibilidad de la señal de voz una vez que se ha aplicado la eliminación de ruido. En la figura 5.7 (b) se presenta el comportamiento de la métrica PESQ con relación a los distintos niveles de SNR. Esta métrica evalúa la calidad perceptual de la señal de voz después de la eliminación de ruido, lo que permite medir el impacto subjetivo de la mejora realizada. Al analizar ambas sub-figuras, se puede observar cómo las métricas STOI y PESQ varían en función de los diferentes niveles de SNR. Estas métricas son indicadores cuantitativos que permiten evaluar la efectividad de la eliminación de ruido en la mejora de la inteligibilidad y la calidad perceptual de la señal de voz.

En la Tabla 5.3 se presentan los resultados cuantitativos de cada métrica evaluada. En el caso específico del ruido de fondo de carro en la señal de voz, se obtiene un valor de 0.7643 para la métrica STOI. Sin embargo, al aplicar la eliminación de ruido utilizando la red neuronal profunda, se logra un promedio de 0.8623, lo que indica una mejora significativa en la inteligibilidad de la señal. Posteriormente, al aplicar la transformada a esta última etapa de procesamiento, se observa una ligera pérdida en la métrica STOI de tan solo 0.0066, la cual se considera irrelevante en términos prácticos. En cuanto a la métrica PESQ, se obtiene un promedio de 1.6643 cuando la señal de habla presenta ruido de fondo. Después de aplicar los procedimientos de eliminación de ruido utilizando únicamente la red neuronal profunda, se alcanza un valor promedio de 2.4497. Al adicionar la técnica de filtrado a esta última etapa de procesamiento, se logra un valor promedio de 2.5777 en la métrica PESQ. Estos resultados demuestran una mejora sustancial en la calidad

perceptual de la señal de voz a medida que se aplican los procesos de eliminación de ruido propuestos.

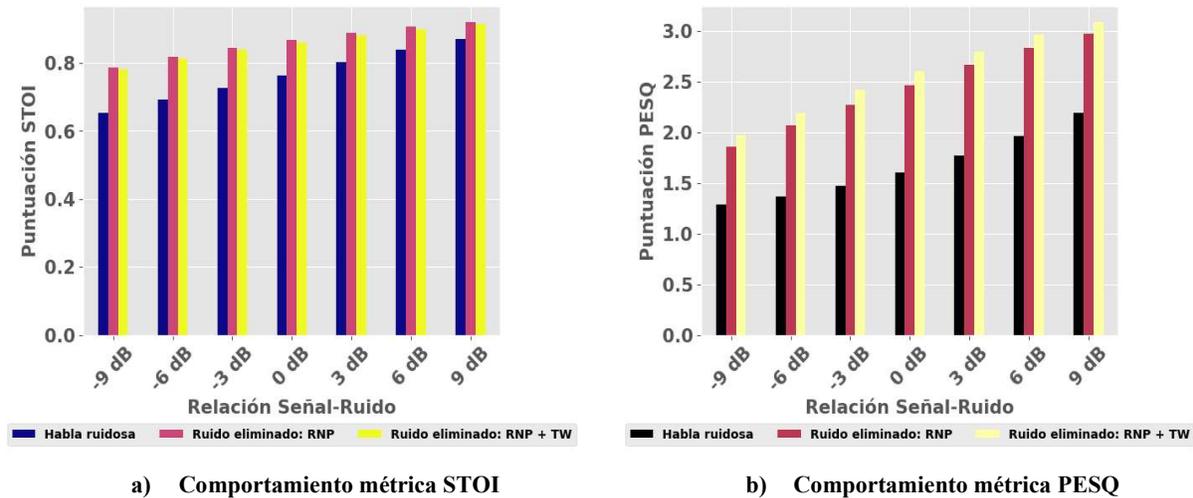


Figura 5.7 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: carro

Tabla 5.3 Puntajes STOI y PESQ para corpus LibreSpeech - Tipo de ruido: Carro

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.6548	0.7859	0.7805	1.2839	1.8616	1.9742
-6 dB	0.6917	0.8181	0.8126	1.3623	2.0665	2.1912
-3 dB	0.7273	0.8451	0.8387	1.4693	2.2752	2.4183
0 dB	0.7646	0.8680	0.8613	1.6037	2.4626	2.6067
3 dB	0.8023	0.8898	0.8825	1.7694	2.6672	2.7967
6 dB	0.8381	0.9076	0.9005	1.9678	2.8373	2.9642
9 dB	0.8715	0.9218	0.9143	2.1939	2.9775	3.0929
Promedio	0.7643	0.8623	0.8557	1.6643	2.4497	2.5777

5.2.4 Resultados para tipo de ruido estacionario – cabina

A continuación, se aborda el caso del tratamiento de eliminación de ruido de cabina. En la figura 5.8 se muestra el comportamiento de dicho proceso para diferentes niveles de SNR. En particular, en la figura 5.8 (a) se presenta el comportamiento de mejora de la métrica de STOI, la cual evalúa la inteligibilidad de la señal. Por otro lado, en la figura 5.8 (b) se muestra el comportamiento de mejora de la métrica de PESQ, que evalúa la calidad perceptual de la señal. Estas figuras proporcionan una representación visual del impacto de la eliminación de ruido de cabina en las métricas STOI y PESQ. El comportamiento de la métrica STOI en la figura 5.8 (a) muestra una clara tendencia de mejora a medida que se reducen los niveles de ruido de cabina, lo cual se refleja en una mayor inteligibilidad de la señal de voz. Por su parte, la figura 5.8 (b) muestra

cómo la métrica de PESQ mejora a medida que disminuye el ruido de cabina, lo que indica una mejora en la calidad perceptual de la señal. Estas representaciones gráficas evidencian el efecto positivo de la eliminación de ruido de cabina en las métricas de evaluación de la señal de voz.

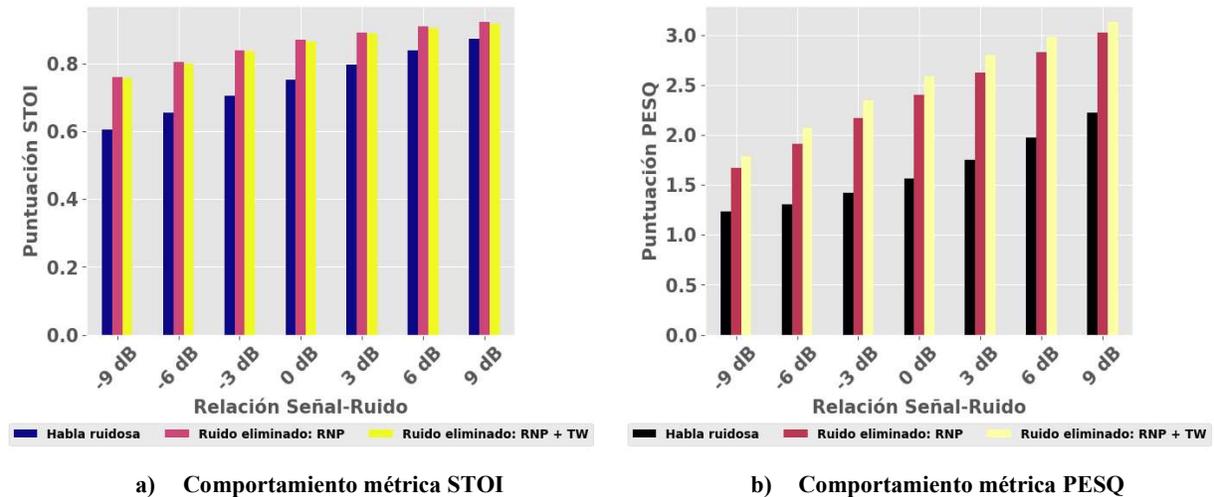


Figura 5.8 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: cabina

En la tabla 5.4 se presentan los resultados numéricos obtenidos para el tratamiento de ruido de cabina en diferentes niveles de SNR, mostrando las métricas de STOI y PESQ respectivamente. Estos datos cuantitativos permiten evaluar el desempeño del sistema en cada escenario analizado. En relación con la métrica de STOI, los resultados muestran que una señal afectada por ruido de cabina tiene un puntaje promedio de 0.7478. Sin embargo, al aplicar la red neuronal propuesta para mejorar la señal, se logra un incremento en el puntaje promedio de STOI de 0.8582. Al realizar el procesamiento adicional mediante la transformada Wavelet, se observa una ligera pérdida de 0.0041 en el puntaje de STOI. Por otro lado, al analizar los puntajes de PESQ en los diferentes escenarios, se encuentra que una señal con ruido de cabina tiene un valor promedio de PESQ de 1.6383. Al aplicar la eliminación de ruido mediante la red neuronal, se logra un incremento significativo en el puntaje de PESQ, alcanzando un valor promedio de 2.3773. Además, al aplicar la transformada a este último procesamiento, se obtiene un puntaje promedio de PESQ de 2.5312. Estos resultados demuestran la capacidad del sistema propuesto para mejorar la calidad perceptual de la señal de voz afectada por ruido de cabina. La tabla 5.4 proporciona una visión cuantitativa de los efectos del tratamiento de ruido de cabina en términos de las métricas STOI y PESQ. Estos resultados respaldan la eficacia de los métodos utilizados en la reducción y eliminación de ruido de cabina, lo que se traduce en una mejora significativa en la inteligibilidad y calidad perceptual de la señal de voz en entornos con presencia de este tipo de ruido.

5.2.5 Resultados para tipo de ruido estacionario – lluvia

A continuación, se realiza un análisis del comportamiento del tratamiento del ruido de lluvia. La figura 5.9 muestra las diferentes mejoras obtenidas al tratar este tipo de ruido. En la figura 5.9 (a) se presenta el comportamiento de la métrica de STOI en función de los distintos niveles de SNR. Por otro lado, la figura 5.9 (b) muestra cómo se mejora la métrica de PESQ al procesar la

señal afectada por este tipo de ruido. En la figura 5.9 (a) se observa cómo la métrica de STOI se ve afectada por el ruido de lluvia en diferentes niveles de SNR. A medida que se aplica el tratamiento propuesto, se evidencia una mejora en la métrica de STOI, lo cual indica una mayor inteligibilidad de la señal de voz. Por su parte, en la figura 5.9 (b) se muestra el comportamiento de la métrica de PESQ al mejorar la señal que contiene ruido de lluvia. Se observa que, a medida que se aplica el tratamiento, la métrica de PESQ experimenta una mejora significativa, lo que se traduce en una mejor calidad perceptual de la señal de voz.

Tabla 5.4 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: cabina

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.6052	0.7606	0.7579	1.2313	1.6740	1.7900
-6 dB	0.6560	0.8048	0.8015	1.3034	1.9106	2.0692
-3 dB	0.7065	0.8412	0.8372	1.4186	2.1698	2.3476
0 dB	0.7543	0.8703	0.8664	1.5679	2.4047	2.5875
3 dB	0.7991	0.8937	0.8894	1.7491	2.6267	2.8018
6 dB	0.8397	0.9116	0.9069	1.9713	2.8328	2.9845
9 dB	0.8742	0.9253	0.9197	2.2267	3.0226	3.1379
Promedio	0.7478	0.8582	0.8541	1.6383	2.3773	2.5312

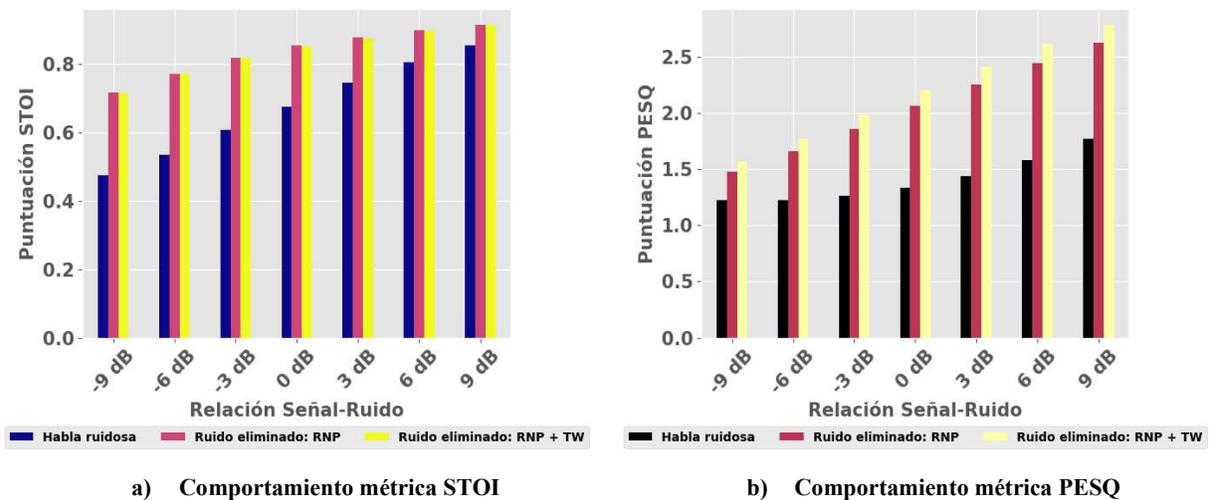


Figura 5.9 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: lluvia

En la tabla 5.5 se presentan los resultados numéricos de las métricas de STOI y PESQ para cada escenario en el tratamiento del ruido de lluvia. Se realiza un análisis detallado de la mejora obtenida en cada métrica para evaluar la efectividad del enfoque propuesto. En cuanto a la métrica de STOI, se observa que cuando la señal contiene ruido de lluvia, el promedio de la métrica es de 0.6703. Sin embargo, al aplicar la red neuronal propuesta para mejorar la señal de habla, se logra un incremento en la métrica de STOI, alcanzando un promedio de 0.8363. Posteriormente, al

aplicar la transformada Wavelet como segundo proceso de tratamiento del ruido, se registra una ligera disminución de 0.0019 en la métrica de STOI. Por otro lado, en relación a la métrica de PESQ, se evidencia que una señal con ruido de lluvia tiene un puntaje promedio de 1.4042. Al aplicar la red neuronal propuesta, se logra una mejora notable en la métrica de PESQ, alcanzando un promedio de 2.0521. Además, al aplicar la técnica de filtrado basada en la transformada Wavelet como segundo proceso, se obtiene una mejora adicional, elevando la métrica de PESQ a un promedio de 2.1915. Estos resultados indican una mejora significativa en la calidad perceptual de las señales de voz afectadas por el ruido de lluvia.

Tabla 5.5 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: lluvia

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.4736	0.7156	0.7151	1.2236	1.4761	1.5645
-6 dB	0.5356	0.7721	0.7713	1.2195	1.6597	1.7715
-3 dB	0.6065	0.8192	0.8175	1.2635	1.8529	1.9865
0 dB	0.6748	0.8540	0.8511	1.3327	2.0605	2.2052
3 dB	0.7446	0.8785	0.8759	1.4374	2.2505	2.4115
6 dB	0.8040	0.8998	0.8971	1.5820	2.4401	2.6155
9 dB	0.8534	0.9154	0.9130	1.7707	2.6252	2.7864
Promedio	0.6703	0.8363	0.8344	1.4042	2.0521	2.1915

5.2.6 Resultados para tipo de ruido estacionario – viento

El gráfico en la figura 5.10 ilustra el comportamiento de una señal de voz que contiene ruido de tipo viento, así como las mejoras obtenidas mediante el uso de la red neuronal profunda y la transformada Wavelet. En la figura 5.10 (a) se representa la métrica de STOI, mientras que en la figura 5.10 (b) se muestra la métrica de PESQ. En la figura 5.10 (a) se observa la mejora de la métrica de STOI en función de los diferentes niveles de SNR. Se puede apreciar cómo la métrica de STOI se incrementa de manera significativa al aplicar la red neuronal profunda y posteriormente la transformada Wavelet. Estos resultados indican que el enfoque propuesto es efectivo para aumentar la inteligibilidad de la señal de voz afectada por el ruido. Por otro lado, en la figura 5.10 (b) se muestra la mejora de la métrica de PESQ en función de los diferentes niveles de SNR. Se puede observar que, al aplicar la red neuronal profunda y la transformada Wavelet, se obtiene una notable mejora en la métrica de PESQ. Esto indica que el enfoque propuesto también logra mejorar la calidad perceptual de la señal de voz, haciéndola más cercana a la señal de referencia libre de ruido.

En la tabla 5.6 se presentan los puntajes numéricos de las métricas STOI y PESQ para diferentes niveles de SNR al tratar el ruido de tipo viento. En cuanto a la métrica STOI, se observa que una señal de voz afectada por el ruido de tipo viento tiene un promedio de 0.7791 en los distintos niveles de SNR. Sin embargo, al aplicar la red neuronal para la eliminación de ruido, se logra una mejora promedio de 0.8843 en la métrica STOI. Es importante destacar que, al aplicar posteriormente la transformada Wavelet, se produce una ligera degradación de la métrica, con una disminución promedio de 0.0054. Estos resultados indican que la red neuronal es efectiva para

mejorar la inteligibilidad de la señal de voz afectada por el ruido de viento, aunque se debe tener en cuenta el impacto adicional de la transformada Wavelet en la calidad final de la señal. Por otro lado, en la métrica de PESQ, se observa que una señal de voz con ruido de viento tiene un puntaje de 1.7598. Sin embargo, al aplicar los procesamientos para eliminar el ruido, se logra un aumento significativo en el puntaje de PESQ, alcanzando 2.5803 y 2.6782 respectivamente, al utilizar únicamente la red neuronal y al combinarla con la transformada Wavelet. Estos resultados demuestran que el enfoque propuesto no solo mejora la inteligibilidad de la señal, sino que también tiene un impacto positivo en la calidad perceptual de la misma.

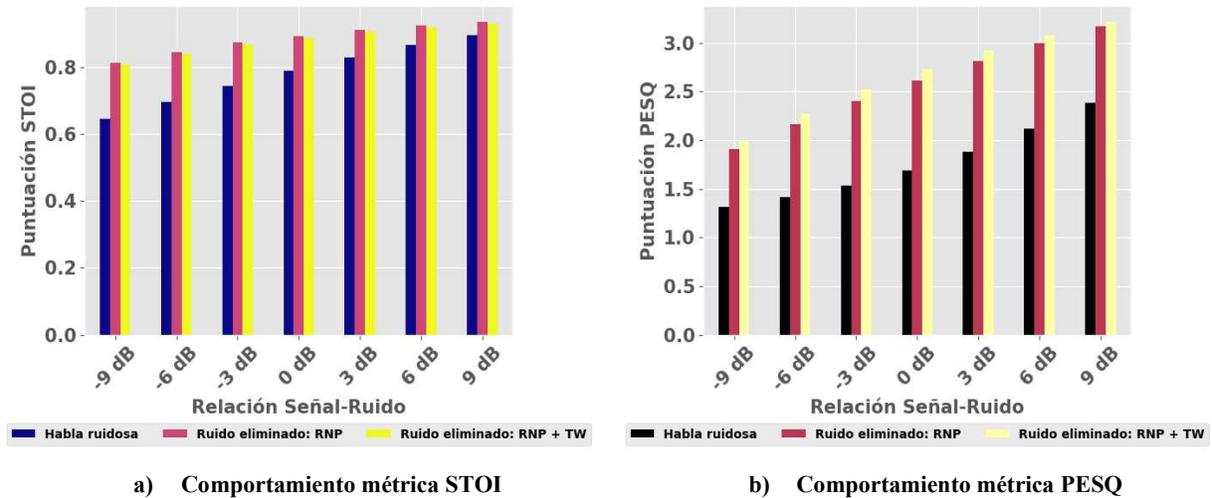


Figura 5.10 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: viento

Tabla 5.6 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: viento

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.6435	0.8120	0.8078	1.3145	1.9061	1.9946
-6 dB	0.6939	0.8442	0.8396	1.4094	2.1612	2.2770
-3 dB	0.7426	0.8721	0.8672	1.5300	2.3997	2.5244
0 dB	0.7873	0.8921	0.8870	1.6882	2.6095	2.7300
3 dB	0.8289	0.9103	0.9044	1.8826	2.8183	2.9216
6 dB	0.8642	0.9245	0.9178	2.1139	2.9981	3.0793
9 dB	0.8934	0.9355	0.9286	2.3805	3.1695	3.2205
Promedio	0.7791	0.8843	0.8789	1.7598	2.5803	2.6782

5.2.7 Resultados para tipo de ruido estacionario – escritura de teclado

Por último, se examina el comportamiento de las métricas al eliminar el ruido de escritura de teclado, utilizando la figura 5.11 como referencia. En la figura 5.11 (a) se muestra el análisis de la métrica de STOI, mientras que en la figura 5.11 (b) se presenta el comportamiento de la métrica

de PESQ. Estas métricas son evaluadas en diferentes niveles de SNR, permitiendo así un análisis exhaustivo de su desempeño en la eliminación del ruido de escritura de teclado.

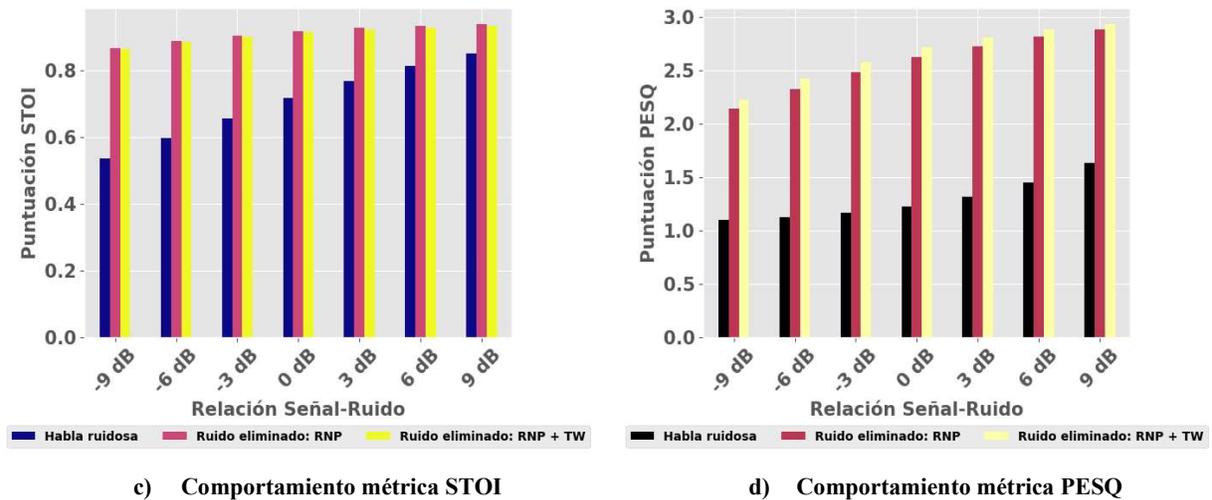


Figura 5.11 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: escritura de teclado

En la tabla 5.7 se presentan los resultados de la evaluación de las métricas subjetivas para la eliminación del ruido de escritura de teclado. Se analiza en primer lugar la métrica de STOI, la cual indica que una señal con este tipo de ruido obtiene un promedio de STOI de 0.7043. Sin embargo, al aplicar la red neuronal, se logra una mejora significativa con un valor promedio de 0.9100. Posteriormente, al aplicar la transformada Wavelet como un segundo proceso, se observa una degradación de la métrica en 0.0040. Por otro lado, la métrica de PESQ indica que una señal con ruido de escritura de teclado obtiene un puntaje de 1.2869. Al aplicar los respectivos procesamientos para mejorar la señal, se logra alcanzar un puntaje de 2.5744 utilizando la red neuronal profunda y 2.6562 al aplicar adicionalmente la transformada Wavelet. Estos resultados demuestran la eficacia de los procedimientos propuestos para mejorar la calidad de la señal al eliminar el ruido de escritura de teclado.

Tabla 5.7 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: escritura de teclado

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.5344	0.8649	0.8629	1.1005	2.1430	2.2259
-6 dB	0.5968	0.8876	0.8848	1.1264	2.3287	2.4293
-3 dB	0.6562	0.9037	0.9000	1.1632	2.4823	2.5810
0 dB	0.7151	0.9165	0.9125	1.2221	2.6284	2.7179
3 dB	0.7665	0.9260	0.9214	1.3169	2.7314	2.8140
6 dB	0.8112	0.9330	0.9278	1.4459	2.8197	2.8856
9 dB	0.8501	0.9389	0.9331	1.6334	2.8873	2.9403
Promedio	0.7043	0.9100	0.9060	1.2869	2.5744	2.6562

5.3 Resultados de eliminación de ruido en la señal de voz del dominio objetivo

En esta sección se presentan y analizan los resultados obtenidos para la eliminación de ruido considerado como no estacionario en el dominio objetivo. Se hace especial énfasis en los resultados reflejados en la figura 5.2 durante el entrenamiento de la red neuronal, en relación con el grado de degradación del error cuadrático medio y su influencia en la reducción del ruido en el dominio objetivo. Además, se destacan los casos en los que no se logró una mitigación adecuada del ruido. Durante el análisis de los resultados, se observó cómo la figura 5.2 proporciona información valiosa sobre el comportamiento del error cuadrático medio durante el entrenamiento de la red neuronal. Estos resultados son indicativos de cómo afecta la disminución del ruido en el dominio objetivo. Se enfatizará particularmente en aquellos casos en los que no se logró una mitigación exitosa del ruido, lo que permite identificar áreas de mejora y futuras investigaciones.

5.3.1 Resultados para tipo de ruido no estacionario – llanto de bebé

En la figura 5.12 se presenta la evaluación de las métricas de STOI y PESQ en diferentes niveles de SNR para el ruido de llanto de un bebé. En la figura 5.12 (a) se observa que la métrica de STOI experimenta una ligera mejora después de aplicar la técnica de aprendizaje profundo y la transformada Wavelet para eliminar el ruido. Por otro lado, en la figura 5.12 (b) se observa un aumento significativo en la métrica de PESQ.

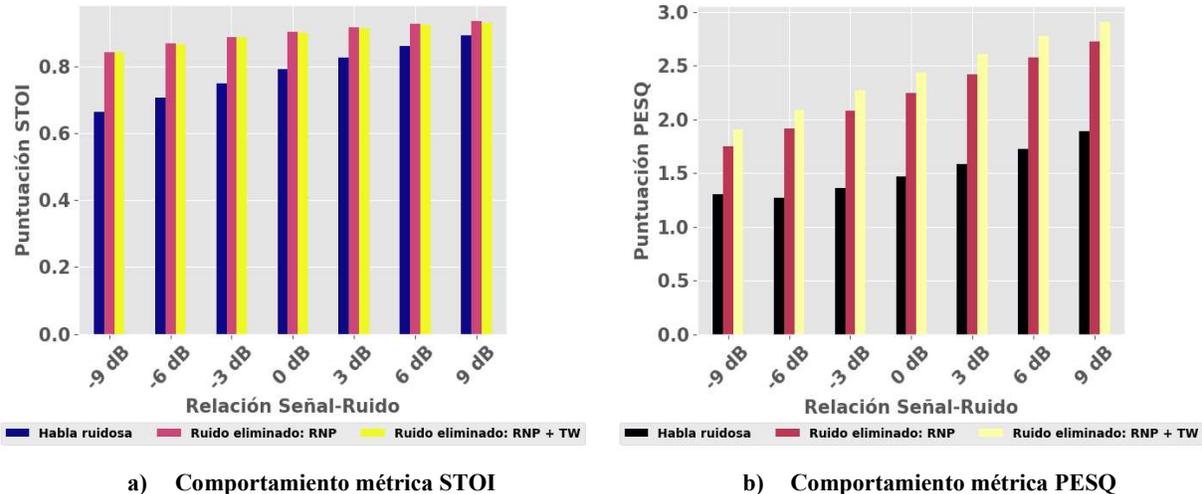


Figura 5.12 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: llanto de bebe

En la tabla 5.8 se presentan los resultados numéricos obtenidos al eliminar el ruido de llanto de un bebé. En primer lugar, se analiza la métrica de STOI, la cual indica que una señal con este tipo de ruido tiene un puntaje promedio de 0.7840. Al aplicar la eliminación de ruido mediante la red neuronal profunda, se logra alcanzar un puntaje promedio de 0.8983. Sin embargo, al aplicar la técnica de filtrado basada en la transformada Wavelet, se observa una ligera disminución en esta métrica de 0.0034. Por otro lado, la métrica de PESQ arroja un puntaje de 1.5138 para una señal con este ruido. Después de mejorar la señal ruidosa mediante la red neuronal, se logra alcanzar un

puntaje promedio de 2.2439. Finalmente, al aplicar la transformada Wavelet a esta última mejora, se obtiene un puntaje promedio de 2.4287.

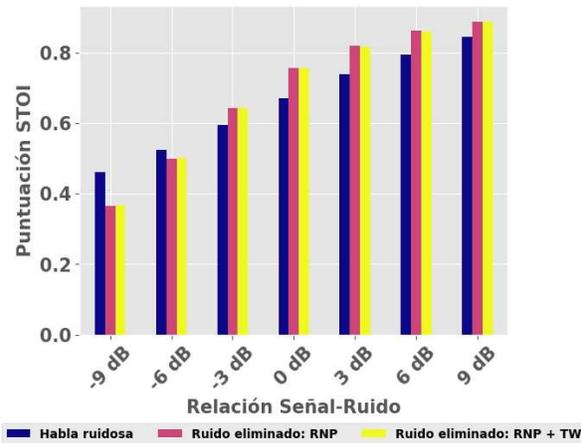
Tabla 5.8 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: llanto de bebe

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.6638	0.8434	0.8419	1.3022	1.7478	1.9061
-6 dB	0.7064	0.8689	0.8672	1.2702	1.9147	2.0911
-3 dB	0.7491	0.8889	0.8866	1.3630	2.0825	2.2676
0 dB	0.7909	0.9044	0.9010	1.4664	2.2430	2.4376
3 dB	0.8258	0.9174	0.9134	1.5834	2.4179	2.6129
6 dB	0.8608	0.9287	0.9235	1.7240	2.5778	2.7754
9 dB	0.8916	0.9367	0.9307	1.8876	2.7242	2.9105
Promedio	0.7840	0.8983	0.8949	1.5138	2.2439	2.4287

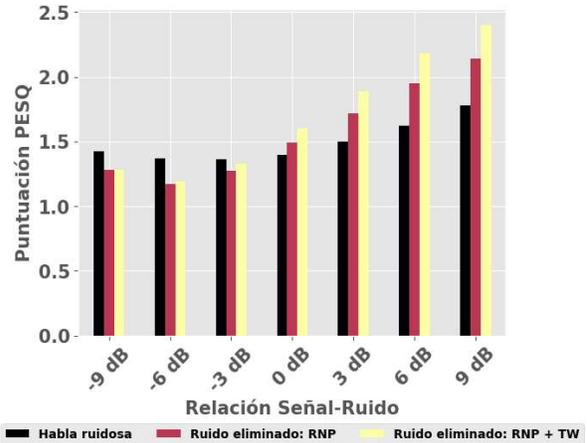
5.3.2 Resultados para tipo de ruido no estacionario – fiesta con multitud de gente

El método propuesto para eliminar el ruido correspondiente a una fiesta con multitud de gente, utilizando la combinación de la red neuronal profunda y la transformada Wavelet, no muestra mejoras en los primeros tres niveles de SNR, como se puede observar en la figura 5.13. Incluso en estos niveles, se indica que la presencia de ruido proporciona una mayor intangibilidad y calidad de la señal. Esto se debe a la naturaleza particular del ruido y al grado de degradación del error cuadrático medio durante el entrenamiento de la red neuronal, lo cual se refleja en las métricas de STOI y PESQ en este caso. En la figura 5.13 (a) se presenta el comportamiento de la métrica de STOI, mientras que en la figura 5.13 (b) se muestra el comportamiento de la métrica del PESQ. Es importante destacar que la eliminación del ruido para este tipo específico presenta una mejora mínima, pero aun así se logra una ligera mejora en la calidad del habla ruidosa de este tipo de ruido.

En la tabla 5.9 se presentan los puntajes numéricos correspondientes a las métricas de STOI y PESQ. Para la métrica de STOI, se realizó la evaluación de la señal que contiene el ruido de una fiesta con multitud de gente, obteniendo un puntaje de 0.6625. Con la mejora mediante la utilización de la red neuronal, se logró un puntaje de 0.6914, y al aplicar la transformada Wavelet se observó una ligera disminución de 0.0002 en comparación con el resultado anterior. En cuanto a la métrica del PESQ, se obtuvo un puntaje de 1.4917 cuando la señal tenía el ruido mencionado. Después de aplicar los métodos propuestos para la mejora, se alcanzaron puntajes de 1.5737 y 1.6964 respectivamente. Es importante destacar que tanto el STOI como el PESQ para este tipo de ruido no experimentaron un aumento significativo como se observó en los ruidos estacionarios, pero aun así se logró una mejora ligeramente notable.



a) Comportamiento métrica STOI



b) Comportamiento métrica PESQ

Figura 5.13 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: fiesta de multitud de gente

Tabla 5.9 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: fiesta de multitud de gente

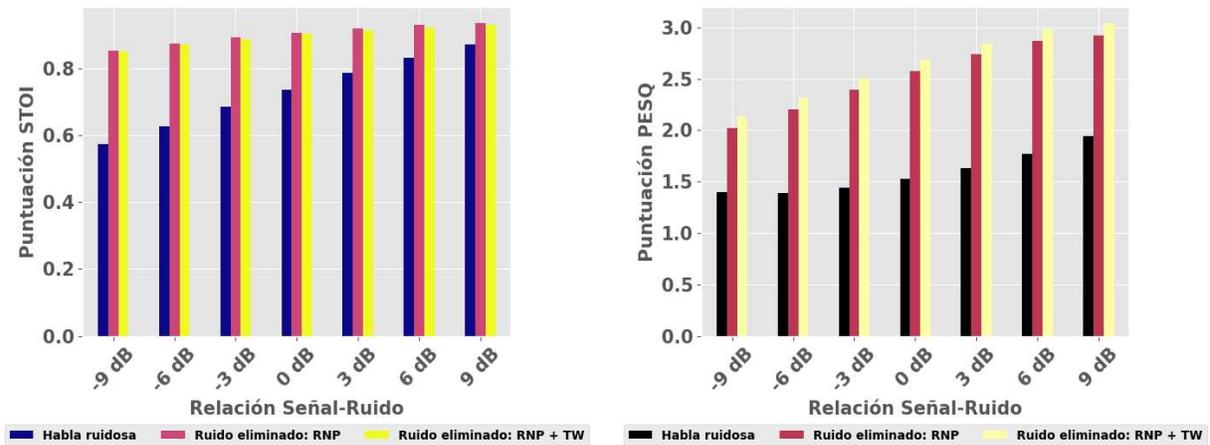
SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.4618	0.3644	0.3689	1.4207	1.2809	1.2769
-6 dB	0.5243	0.4990	0.5024	1.3697	1.1707	1.1895
-3 dB	0.5968	0.6428	0.6434	1.3602	1.2725	1.3309
0 dB	0.6726	0.7586	0.7565	1.3970	1.4883	1.6029
3 dB	0.7400	0.8213	0.8182	1.4972	1.7170	1.8899
6 dB	0.7966	0.8633	0.8607	1.6189	1.9481	2.1826
9 dB	0.8457	0.8904	0.8883	1.7788	2.1389	2.4027
Promedio	0.6625	0.6914	0.6912	1.4917	1.5737	1.6964

5.3.3 Resultados para tipo de ruido no estacionario – campanas de iglesia

Los resultados mostrados en la figura 5.14 representan el comportamiento de los puntajes obtenidos para las métricas de STOI y PESQ al tratar el ruido de campanas de iglesia en diferentes niveles de SNR. En la figura 5.14 (a) se muestra el comportamiento de la métrica de STOI, donde se observa una mejora en los puntajes después de aplicar el procesamiento de eliminación de ruido mediante la red neuronal y la transformada Wavelet. Por otro lado, en la figura 5.14 (b) se presenta la métrica de PESQ, que también muestra una mejora en la calidad del habla para este tipo de ruido.

En la tabla 5.10 se presentan los puntajes correspondientes a las métricas de STOI y PESQ en diferentes niveles de SNR. Comenzando con la métrica de STOI, se observa que, en promedio, una señal contaminada con ruido obtiene un puntaje de 0.7296. Sin embargo, al aplicar técnicas

de eliminación de ruido, como el uso de la red neuronal, se logra mejorar este puntaje a 0.9009. Por otro lado, al aplicar la transformada Wavelet como segundo procesamiento, se experimenta una degradación de la métrica en 0.0046. En cuanto a los puntajes correspondientes al PESQ, se obtiene un valor promedio de 1.5863 para una señal con ruido. Sin embargo, al realizar la eliminación del ruido mediante la red neuronal, se logra una mejora significativa, alcanzando un puntaje de 2.5313. Finalmente, al aplicar el segundo procesamiento con la transformada Wavelet, se alcanza un puntaje de 2.6426.



a) Comportamiento métrica STOI

b) Comportamiento métrica PESQ

Figura 5.14 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: campanas de iglesia

Tabla 5.10 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: campanas de iglesia

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.5740	0.8523	0.8498	1.3955	2.0231	2.1288
-6 dB	0.6260	0.8734	0.8704	1.3909	2.2017	2.3106
-3 dB	0.6839	0.8915	0.8880	1.4405	2.3952	2.5039
0 dB	0.7352	0.9058	0.9014	1.5279	2.5767	2.6862
3 dB	0.7860	0.9190	0.9135	1.6341	2.7369	2.8450
6 dB	0.8317	0.9286	0.9224	1.7698	2.8643	2.9790
9 dB	0.8705	0.9358	0.9288	1.9458	2.9217	3.0450
Promedio	0.7296	0.9009	0.8963	1.5863	2.5313	2.6426

5.3.4 Resultados para tipo de ruido no estacionario – mormullos en cafetería

En esta sección se presentan los resultados del tratamiento del ruido de mormullos en una cafetería. En la figura 5.15 se muestra el comportamiento de las métricas STOI y PESQ en

diferentes niveles de SNR. En la figura 5.15 (a), que corresponde a la métrica STOI, es importante mencionar que el nivel de SNR de -9 dB no experimentó una mejora significativa al procesar la señal con la red neuronal y la transformada Wavelet. Sin embargo, para los demás niveles de SNR, se observa una mejora en esta métrica, como se muestra en el gráfico. Por otro lado, en la figura 5.15 (b) se muestra la evaluación de la métrica PESQ. Aquí se presentan dos casos en los que los niveles de SNR de -9 dB y -6 dB no mostraron una mejora en la eliminación del ruido. De hecho, esta métrica indica que la calidad de la señal de voz es mejor cuando tiene ruido en esos casos particulares. Para los demás niveles de SNR, se observa una mejora en la métrica PESQ.

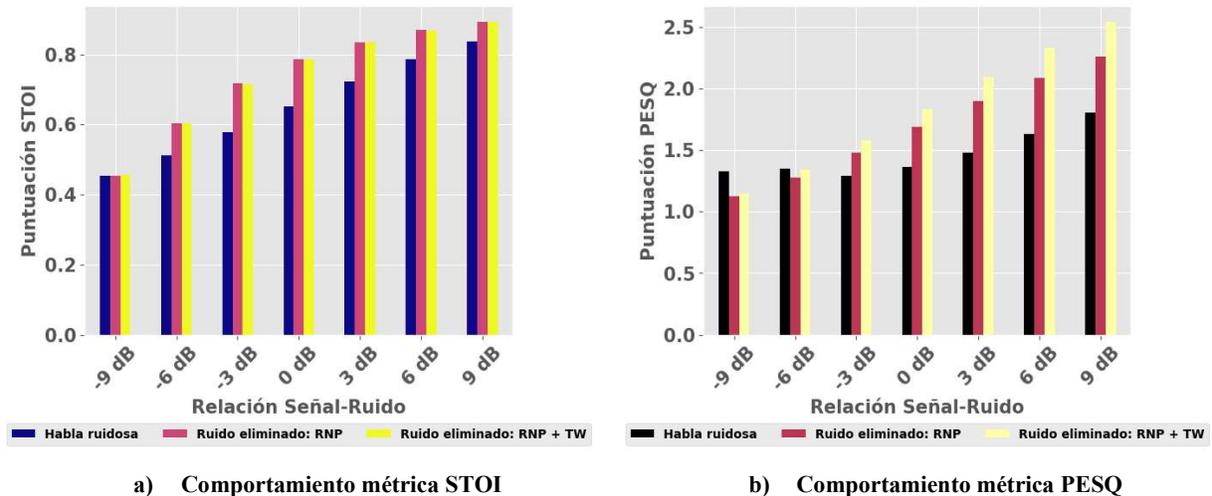


Figura 5.15 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: murmullos de cafetería

En la tabla 5.11 se presentan los resultados numéricos de las puntuaciones de STOI y PESQ al tratar el ruido de murmullos en una cafetería. En general, cuando una señal tiene este tipo de ruido, el puntaje promedio de STOI es de 0.6490. Posteriormente, al aplicar la eliminación de ruido utilizando la red neuronal, se obtiene un puntaje de 0.7370, aunque se pierde un ligero grado de intangibilidad de 0.0004 al someter la señal a un segundo procesamiento con la transformada Wavelet. Por otro lado, la métrica de PESQ muestra un puntaje promedio de 1.4645 cuando la señal contiene ruido. Sin embargo, al aplicar la mejora utilizando la red neuronal, se alcanza un puntaje de 1.6904, y posteriormente, al someter la señal a un segundo procesamiento con la transformada Wavelet, se obtiene finalmente un puntaje de 1.8414.

Tabla 5.11 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: mormullos de cafetería

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.4536	0.4534	0.4569	1.3285	1.1270	1.1515
-6 dB	0.5129	0.6037	0.6041	1.3479	1.2810	1.3402
-3 dB	0.5783	0.7170	0.7163	1.2935	1.4776	1.5820
0 dB	0.6509	0.7869	0.7861	1.3686	1.6934	1.8372
3 dB	0.7241	0.8346	0.8336	1.4789	1.8987	2.0954
6 dB	0.7858	0.8693	0.8674	1.6295	2.0921	2.3353
9 dB	0.8378	0.8943	0.8920	1.8050	2.2632	2.5487
Promedio	0.6490	0.7370	0.7366	1.4645	1.6904	1.8414

5.3.5 Resultados para tipo de ruido no estacionario – helicóptero

En la figura 5.16 se puede observar una mejora gradual en las métricas de STOI y PESQ para la eliminación de ruido de helicóptero mediante las técnicas propuestas. En la figura 5.16 (a) se muestra el comportamiento de la métrica STOI, donde se evidencia una mejora para cada nivel de SNR. Por otro lado, en la figura 5.16 (b) se presentan los resultados de la métrica PESQ, la cual también muestra una mejora en todos los niveles de SNR.

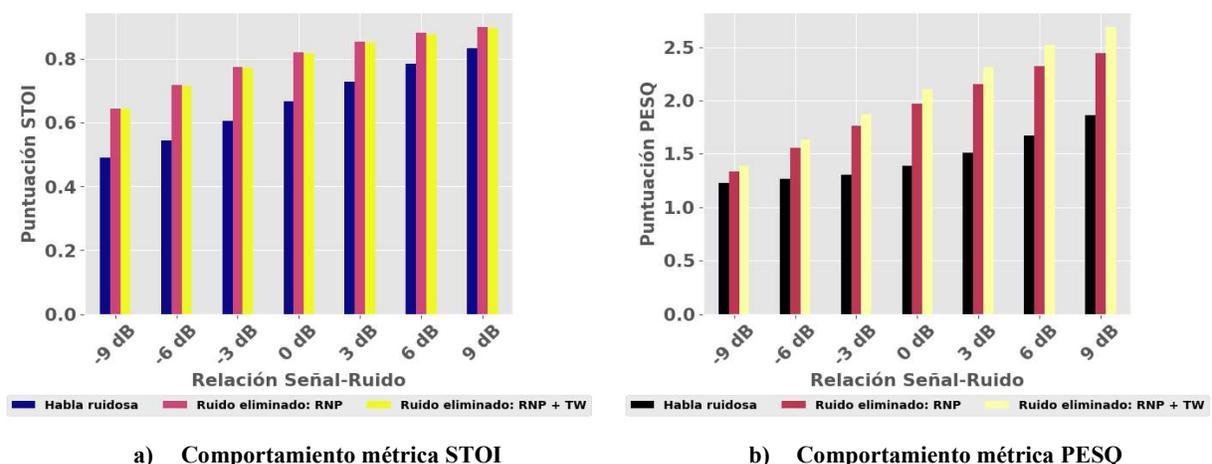


Figura 5.16 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: helicóptero

En la tabla 5.12 se presentan los puntajes generales obtenidos de manera numérica para las métricas subjetivas de STOI y PESQ, que evalúan la calidad de la señal. En primer lugar, se informan los resultados para la métrica STOI. Para una señal con presencia de ruido, el puntaje de STOI es de 0.6652. Sin embargo, al aplicar las técnicas propuestas, el procesamiento a través de la red neuronal mejora este puntaje en un 0.7998. Por otro lado, al emplear la transformada Wavelet como segundo procesamiento para la eliminación del ruido, se observa una ligera pérdida de calidad de la señal de 0.0024 en términos de intangibilidad. En cuanto a la métrica PESQ, en promedio, una señal ruidosa de este tipo obtiene un puntaje de 1.4607. No obstante, al aplicar el

procesamiento para mejorar y eliminar el ruido a través de la red neuronal, se alcanza una puntuación de 1.9346. Finalmente, al emplear la transformada Wavelet, se logra un puntaje de PESQ de 2.0740.

5.3.6 Resultados para tipo de ruido no estacionario – personas hablando

En la figura 5.17 se presentan las métricas de evaluación para la eliminación de ruido en grabaciones de personas hablando. En la figura 5.17 (a) se muestra el comportamiento de la métrica STOI en los distintos niveles de SNR. En la figura 5.17 (b) se presenta el comportamiento de la métrica PESQ. En ambos casos, se observa una mejora ligera en cada uno de los niveles de SNR, indicando una reducción del ruido en las grabaciones de personas hablando.

Tabla 5.12 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: helicóptero

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.4907	0.6454	0.6437	1.2268	1.3379	1.3899
-6 dB	0.5435	0.7197	0.7176	1.2667	1.5567	1.6328
-3 dB	0.6068	0.7751	0.7734	1.3007	1.7653	1.8680
0 dB	0.6686	0.8203	0.8182	1.3917	1.9685	2.1041
3 dB	0.7281	0.8544	0.8523	1.5100	2.1515	2.3153
6 dB	0.7847	0.8817	0.8788	1.6705	2.3193	2.5167
9 dB	0.8343	0.9020	0.8982	1.8587	2.4434	2.6918
Promedio	0.6652	0.7998	0.7974	1.4607	1.9346	2.0740

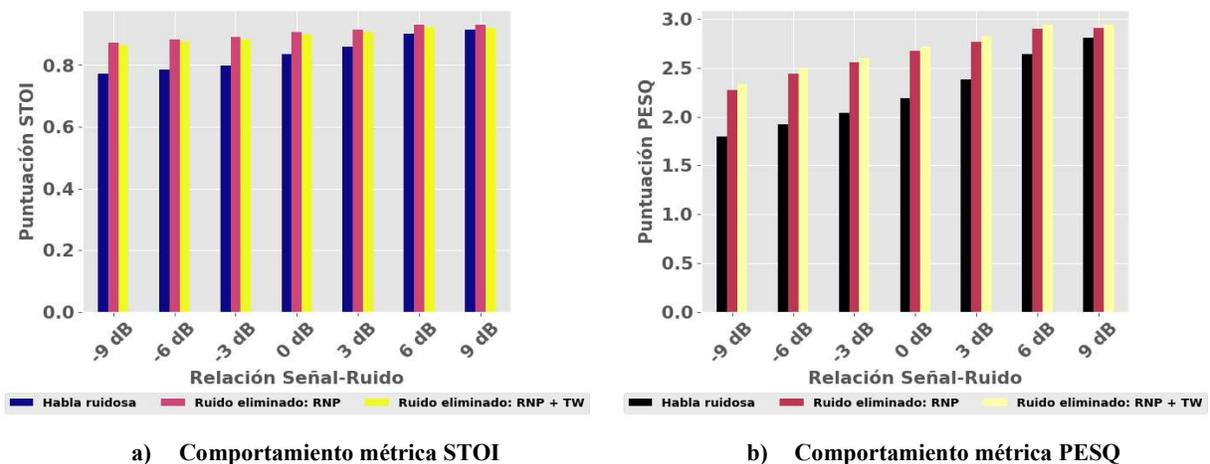


Figura 5.17 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: personas hablando

En la tabla 5.13 se presentan los puntajes de las métricas STOI y PESQ. Al analizar la métrica STOI, se observa que una señal de voz contaminada con ruido de personas hablando tiene, en

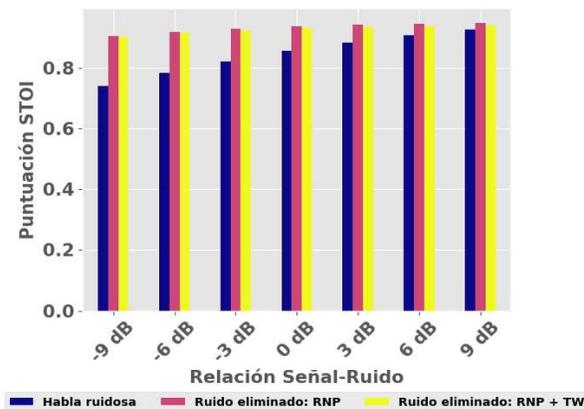
promedio, una evaluación de 0.8387. Sin embargo, al aplicar el procesamiento correspondiente para eliminar dicho ruido utilizando la red neuronal, este puntaje aumenta en 0.9048. En contraste, al aplicar la transformada Wavelet, se registra una disminución de 0.0661 en esta métrica. Por otro lado, al analizar los resultados de la métrica PESQ, se observa que una señal con ruido inicial tiene un puntaje de 2.2542. Al aplicar las técnicas mencionadas para eliminar el ruido, se alcanzan puntajes de 2.6457 y 2.6918, respectivamente. Estos resultados indican una mejora en la calidad de la señal al eliminar el ruido de personas hablando, tanto en términos de la métrica STOI como en la métrica PESQ.

5.3.7 Resultados para tipo de ruido no estacionario – ladrido de perro

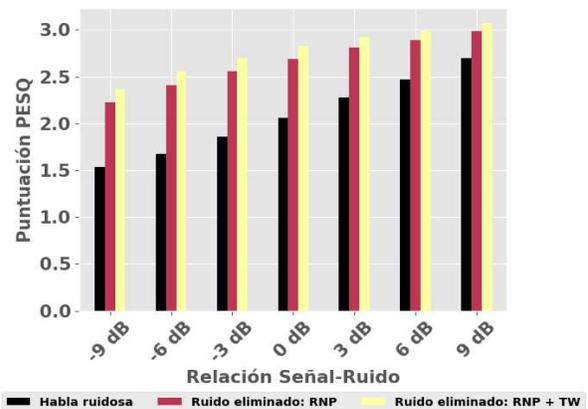
Por último, en la figura 5.18 se exhibe el comportamiento de las métricas para la eliminación del ruido de ladridos de perro. En la figura 5.18 (a) se presentan los puntajes obtenidos para la métrica STOI, mientras que en la figura 5.18 (b) se muestran los puntajes obtenidos para la métrica PESQ. Ambas métricas evidencian una mejora gradual a medida que se incrementan los niveles de SNR.

Tabla 5.13 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: personas hablando

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.7723	0.8722	0.8638	1.7960	2.2721	2.3296
-6 dB	0.7866	0.8844	0.8767	1.9188	2.4362	2.4861
-3 dB	0.7979	0.8915	0.8830	2.0397	2.5550	2.5958
0 dB	0.8360	0.9080	0.8992	2.1928	2.6782	2.7192
3 dB	0.8596	0.9147	0.9070	2.3823	2.7703	2.8240
6 dB	0.9027	0.9322	0.9238	2.6385	2.8992	2.9457
9 dB	0.9162	0.9312	0.9217	2.8116	2.9089	2.9427
Promedio	0.8387	0.9048	0.8964	2.2542	2.6457	2.6918



a) Comportamiento métrica STOI



b) Comportamiento métrica PESQ

Figura 5.18 Puntajes STOI y PESQ en diferentes niveles SNR - tipo de ruido: ladrido de perro

En la tabla 5.14 se presentan los resultados obtenidos de la evaluación subjetiva para la eliminación de ruido. En primer lugar, se analiza la métrica STOI, la cual arroja un promedio de 0.8457 cuando la señal contiene ruido. Posteriormente, al aplicar los procesamientos para la eliminación del ruido, se obtiene un puntaje de 0.9315 al utilizar la red neuronal. Sin embargo, al aplicar la transformada Wavelet, se observa una ligera pérdida de 0.0056 en esta métrica. Por otro lado, se analiza el puntaje obtenido para la métrica PESQ, la cual indica que una señal con ruido tiene un puntaje de 2.0803. Al aplicar las respectivas técnicas de eliminación de ruido, se obtiene una mejora con un puntaje de 2.6510 al utilizar la red neuronal y de 2.7767 al emplear la transformada Wavelet.

Tabla 5.14 Puntajes STOI y PESQ para corpus LibreSpeech - tipo de ruido: ladrido de perro

SNR	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW	Señal ruidosa	Aplicando RNP	Aplicando RNP+TW
	STOI			PESQ		
-9 dB	0.7413	0.9042	0.9014	1.5308	2.2251	2.3628
-6 dB	0.7840	0.9189	0.9151	1.6710	2.4089	2.5553
-3 dB	0.8214	0.9281	0.9232	1.8553	2.5561	2.6980
0 dB	0.8562	0.9358	0.9302	2.0599	2.6850	2.8266
3 dB	0.8840	0.9409	0.9341	2.2814	2.8123	2.9267
6 dB	0.9066	0.9443	0.9371	2.4695	2.8858	2.9901
9 dB	0.9264	0.9487	0.9406	2.6945	2.9842	3.0777
Promedio	0.8457	0.9315	0.9259	2.0803	2.6510	2.7767

Los resultados obtenidos en esta sección de mejora del reconocimiento del habla se analizarán tanto en el contexto de un sistema reconocedor automático de voz como en la identificación de locutor de texto independiente, los cuales serán discutidos en las secciones 5.4 y 5.5 de este capítulo.

5.4 Reconocimiento automático de voz

En esta sección se realiza un análisis de las mejoras obtenidas en el reconocimiento automático de voz para el habla ruidosa. Se presentan las tasas de error de palabra correspondientes a los 14 tipos de ruido tratados tanto en el dominio fuente como en el dominio objetivo. Es importante destacar que la tasa de error de palabra base para los audios sin ruido es de 0.0232.

5.4.1 Resultados para datos fuente

En esta sección se analiza los resultados del dominio fuente.

5.4.1.1 Tipo de ruido estacionario – ruido rosa

En la figura 5.19 se muestra el comportamiento de la tasa de error de palabra al reducir el ruido

rosa. En dicha figura, la línea en color rojo representa el audio con este tipo de ruido, y se puede observar una diferencia significativa en comparación con la línea verde, que representa la tasa de error de palabra base. Mediante las técnicas de eliminación de ruido, se logra disminuir el error, lo cual se refleja en la línea gris que representa solo la aplicación de la red neuronal. Asimismo, la combinación de la red neuronal y la transformada Wavelet, representada por la línea en color, logra una ligera disminución adicional en el error de palabras.

A continuación, se presentan los resultados numéricos de la tasa de error de palabra para abordar el ruido rosa en la tabla 5.15. La tasa de error de palabra promedio para los audios con este tipo de ruido fue de 0.4509. Al eliminar el ruido de la señal de voz, se logró reducir la tasa de error de palabra a 0.2203, lo que representa una mejora adicional de 0.0356 en la disminución de la tasa de error de palabra.

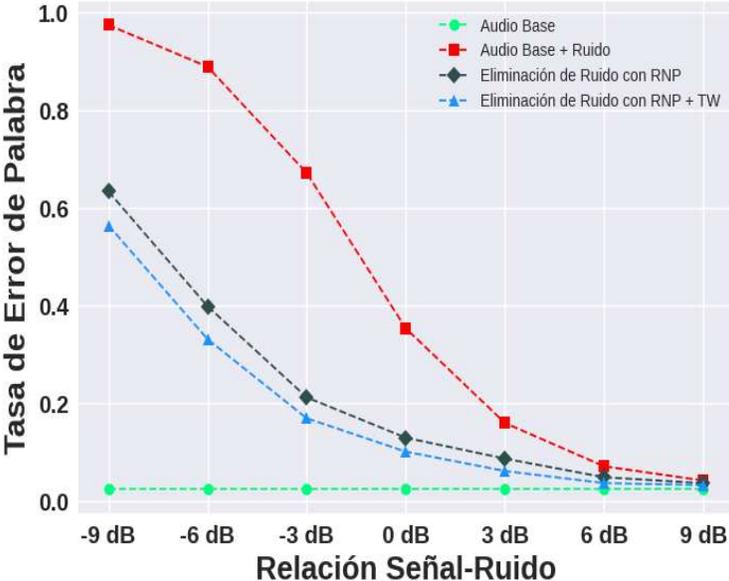


Figura 5.19 Comportamiento de la tasa de error de palabra - tipo de ruido: rosa

Tabla 5.15 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: rosa

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.9737	0.6331	0.5638
-6 dB	0.8883	0.3980	0.3308
-3 dB	0.6709	0.2117	0.1688
0 dB	0.3528	0.1290	0.1005
3 dB	0.1591	0.0863	0.0609
6 dB	0.0705	0.0482	0.0362
9 dB	0.0415	0.0359	0.0322
Promedio	0.4509	0.2203	0.1847

5.4.1.2 Tipo de ruido estacionario – gotas de agua

En el gráfico de la figura 5.20 se puede observar el comportamiento de la tasa de error de

palabra en presencia de habla ruidosa (línea en color rojo). Asimismo, se muestra el resultado obtenido al eliminar el ruido utilizando una red neuronal profunda (línea en color gris), y posteriormente, se aplica la transformada Wavelet a dicho proceso mencionado. El comportamiento se analiza en función de los diferentes niveles de SNR, y se puede apreciar cómo las mejoras logran disminuir de manera efectiva el ruido de las gotas de agua.

En la tabla 5.16 se presentan los resultados numéricos de la tasa de error de palabra al tratar el ruido de gotas de agua en distintos niveles de SNR. En general, se destaca que, al enfrentar este tipo de ruido, se obtiene un promedio acumulado de 0.1040 en la tasa de error de palabra. Sin embargo, al mitigar este ruido mediante el uso de la red neuronal, se logra reducir la tasa de error de palabra a 0.0327. Además, al aplicar la transformada Wavelet a este proceso de mitigación, se consigue disminuir aún más el error a 0.0297.

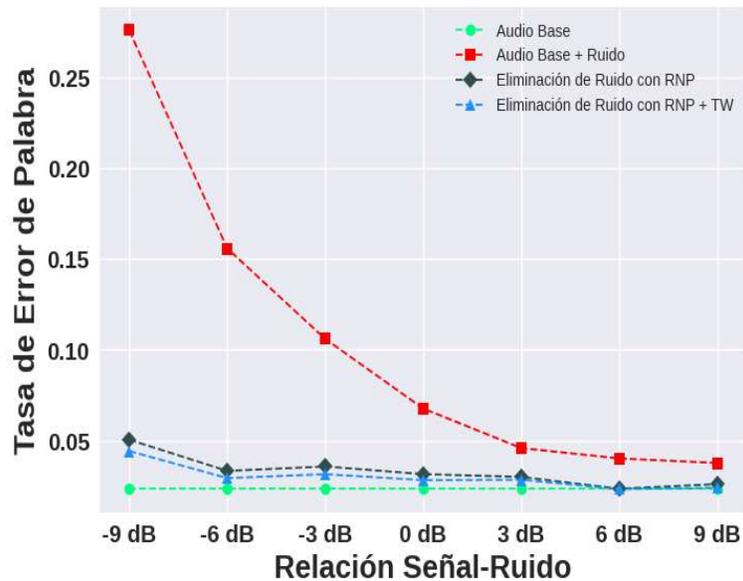


Figura 5.20 Comportamiento de la tasa de error de palabra - tipo de ruido: gotas de agua

Tabla 5.16 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: gotas de agua

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.2758	0.0503	0.0441
-6 dB	0.1557	0.0331	0.0292
-3 dB	0.1058	0.0356	0.0313
0 dB	0.0675	0.0313	0.0281
3 dB	0.0457	0.0298	0.0283
6 dB	0.0400	0.0233	0.0231
9 dB	0.0375	0.0259	0.0239
Promedio	0.1040	0.0327	0.0297

5.4.1.3 Tipo de ruido estacionario – carro

En el gráfico de la figura 5.21 se presenta la tasa de error de palabra para cada situación

relacionada con el ruido de carro. Es importante destacar que, a pesar de obtener mejoras en las métricas de STOI y PESQ al tratar este tipo de ruido, al aplicar la red neuronal y la transformada para la eliminación de ruido, la tasa de error de palabra no disminuye, e incluso la línea en rojo, que representa los audios con ruido, muestra un rendimiento superior en comparación con la aplicación de las técnicas de eliminación de ruido. Esto se debe a la propia naturaleza del ruido, además de que durante los procesamientos se puede perder intangibilidad y calidad de la señal. Aunque las diferencias entre las líneas gris y azul son mínimas, el reconocedor automático de voz presenta un mejor rendimiento con el ruido incrustado en los diferentes niveles de SNR.

En la tabla 5.17 se presentan los resultados numéricos para el caso de eliminación de ruido de carro. Para una señal de voz con este tipo de ruido, se obtiene una tasa promedio de error de palabra de 0.0456. Sin embargo, al aplicar la red neuronal para la eliminación de ruido, se observa un ligero aumento en la tasa de error de palabra, alcanzando 0.0575. Posteriormente, al aplicar la transformada Wavelet a este método, se logra obtener una tasa promedio de error de 0.0508. Aunque se logra una ligera mejora, no es suficiente para alcanzar el valor de 0.0456 obtenido cuando el audio original contiene el ruido de carro.

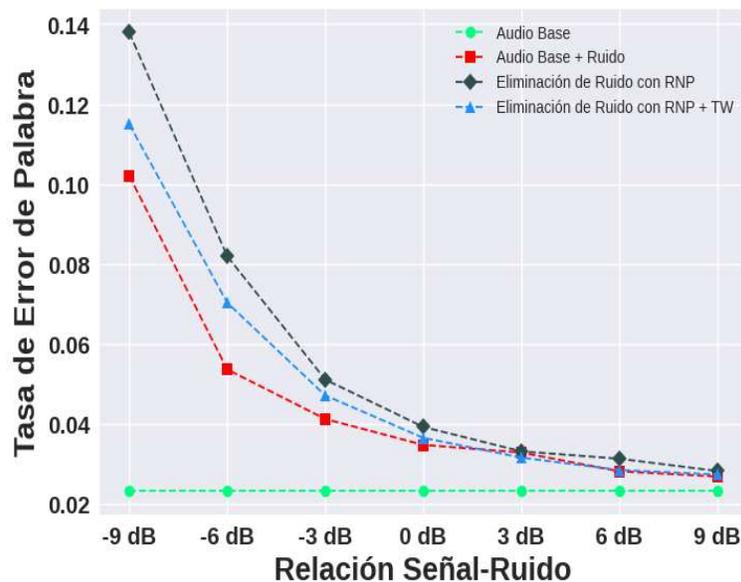


Figura 5.21 Comportamiento de la tasa de error de palabra - tipo de ruido: carro

Tabla 5.17 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: carro

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.1020	0.1378	0.1149
-6 dB	0.0536	0.0819	0.0703
-3 dB	0.0412	0.0511	0.0471
0 dB	0.0347	0.0392	0.0365
3 dB	0.0328	0.0331	0.0316
6 dB	0.0282	0.0313	0.0285
9 dB	0.0269	0.0282	0.0273
Promedio	0.0456	0.0575	0.0508

5.4.1.4 Tipo de ruido estacionario – cabina

En la figura 5.22 se presenta un caso particular del comportamiento de la tasa de error de palabra al tratar el ruido de cabina. En el gráfico se puede observar que la línea roja representa la tasa de error de palabra cuando el audio contiene dicho ruido. Sin embargo, la reducción de este ruido utilizando las técnicas propuestas es mínima, como se puede apreciar en las líneas en color gris y azul. Aunque se logró una mejora, no fue significativa en comparación con los otros tipos de ruidos estacionarios abordados en este estudio.

En la tabla 5.18 se confirma lo mencionado en la figura anterior. De acuerdo con los datos obtenidos, la tasa de error de palabra cuando el habla está afectada por el ruido es de 0.1097. Se observa una ligera mejora que no resulta significativa cuando se aplica la red neuronal para mejorar el habla, reduciendo el error a 0.1045. Finalmente, al aplicar la transformada Wavelet para la eliminación de ruido a esta última técnica, se obtiene un error de 0.0919. En ambas técnicas de eliminación de ruido, no se logró obtener una mejora significativa en la tasa de error de palabra.

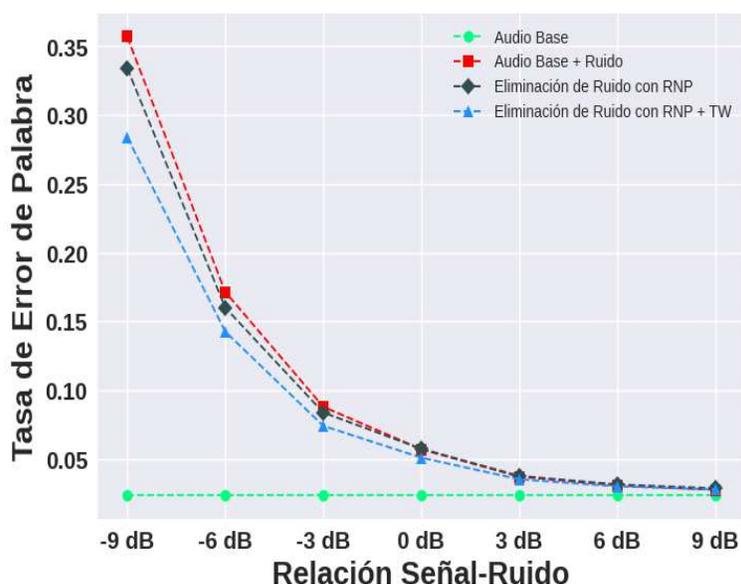


Figura 5.22 Comportamiento de la tasa de error de palabra - tipo de ruido: cabina

Tabla 5.18 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: cabina

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.3569	0.3341	0.2839
-6 dB	0.1714	0.1599	0.1425
-3 dB	0.0878	0.0836	0.0739
0 dB	0.0569	0.0572	0.0508
3 dB	0.0369	0.0375	0.0350
6 dB	0.0306	0.0313	0.0299
9 dB	0.0274	0.0285	0.0276
Promedio	0.1097	0.1045	0.0919

5.4.1.5 Tipo de ruido estacionario – lluvia

En la figura 5.23 se muestra la tasa de error de palabra al mitigar el ruido estacionario tipo lluvia. En el gráfico se representan cuatro líneas de colores distintos. La línea verde corresponde a la tasa de error de palabra cuando los audios no contienen ruido. Por otro lado, la línea roja representa la tasa de error de palabra cuando se añade el ruido de lluvia a los audios, mostrando un alto nivel de error en comparación con la línea base. En color gris, se muestra la mejora en el reconocimiento de voz mediante el uso de la red neuronal, mientras que la línea azul representa la implementación conjunta de la red neuronal y la transformada Wavelet. Estas técnicas implementadas logran mejoras en comparación con los audios afectados por el ruido en los diferentes niveles de SNR.

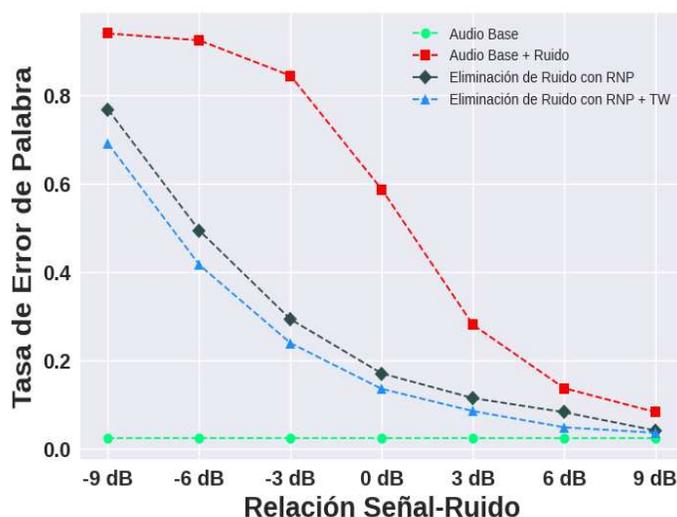


Figura 5.23 Comportamiento de la tasa de error de palabra - tipo de ruido: lluvia

En la tabla 5.19 se presentan los resultados obtenidos de la tasa de error de palabra para cada uno de los escenarios tratados en la eliminación de ruido de tipo específico. El primer caso corresponde al audio con ruido, el cual muestra una tasa de error de palabra promedio de 0.5421. A continuación, se muestra el caso en el que se aplica la eliminación de ruido mediante la red neuronal, logrando una disminución en la tasa de error de palabra de 0.2805. Por último, se presenta el resultado obtenido al aplicar la combinación de la red neuronal y la transformada Wavelet para la eliminación de ruido, evidenciando una mejora en el desempeño con un error de 0.2359.

Tabla 5.19 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: lluvia

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.9395	0.7675	0.6905
-6 dB	0.9239	0.4938	0.4161
-3 dB	0.8439	0.2931	0.2391
0 dB	0.5868	0.1705	0.1356
3 dB	0.2804	0.1147	0.0853
6 dB	0.1371	0.0828	0.0486
9 dB	0.0832	0.0414	0.0364
Promedio	0.5421	0.2805	0.2359

5.4.1.6 Tipo de ruido estacionario – viento

En la figura 5.24 se presenta el comportamiento de la tasa de error de palabra en diferentes niveles de SNR para el tratamiento de eliminación de ruido de tipo viento. La línea roja muestra la tasa de error de palabra cuando el audio del habla está contaminado con este tipo de ruido. Por otro lado, las líneas gris y azul representan las técnicas implementadas para eliminar el ruido de tipo viento. Se puede observar una mejora en la tasa de error de palabra al aplicar estas técnicas de eliminación de ruido. Aunque la mejora es moderada, se obtiene una reducción del error en la tasa de error de palabra.

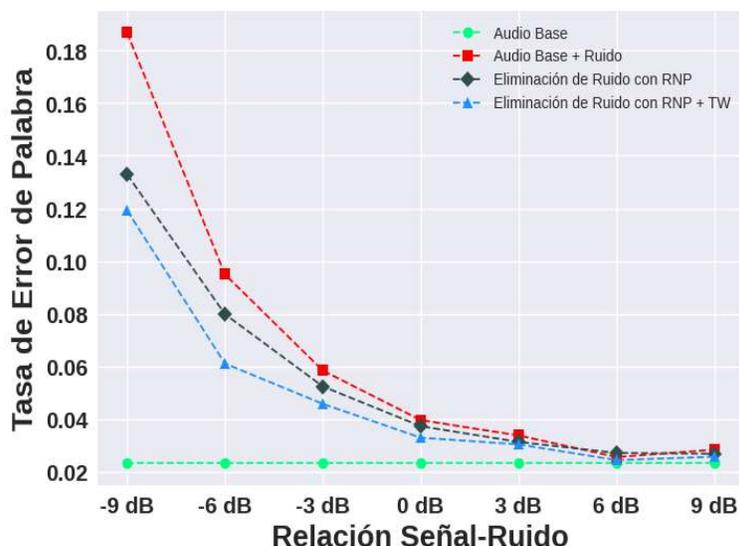


Figura 5.24 Comportamiento de la tasa de error de palabra - tipo de ruido: viento

En la tabla 5.20 se presentan los resultados de la tasa de error de palabra para la eliminación de ruido de tipo viento en diferentes niveles de SNR. En primer lugar, se muestra la tasa de error de palabra promedio cuando los audios del habla están contaminados con este tipo de ruido, con un valor de 0.0667. Luego, al aplicar la red neuronal profunda, se logra una mejora promedio de 0.0553 en la tasa de error de palabra. Finalmente, al combinar este último procesamiento con la transformada Wavelet, se alcanza una ligera mejora adicional, con una tasa de error de palabra de 0.0485.

Tabla 5.20 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: viento

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.1867	0.1329	0.1192
-6 dB	0.0949	0.0798	0.0611
-3 dB	0.0584	0.0524	0.0458
0 dB	0.0396	0.0373	0.0329
3 dB	0.0338	0.0313	0.0304
6 dB	0.0256	0.0271	0.0244
9 dB	0.0284	0.0268	0.0257
Promedio	0.0667	0.0553	0.0485

5.4.1.7 Tipo de ruido estacionario – escritura de teclado

En el dominio fuente, también se presentó el ruido estacionario de tipo escritura de teclado. En la figura 5.25 se muestra el comportamiento de la tasa de error de palabra al tratar de eliminar este tipo de ruido. Las líneas en color gris y azul representan la reducción de la tasa de error de palabra utilizando las técnicas propuestas en este trabajo. En este caso, se logró mitigar en gran medida el ruido en diferentes niveles de SNR. Es importante destacar que la aplicación de la transformada Wavelet como segundo procesamiento no resultó necesaria en su totalidad, ya que la disminución del error fue ligera en los primeros niveles de SNR. En los demás niveles de SNR, la aplicación de la transformada Wavelet tuvo un efecto similar a la eliminación de ruido mediante la red neuronal profunda.

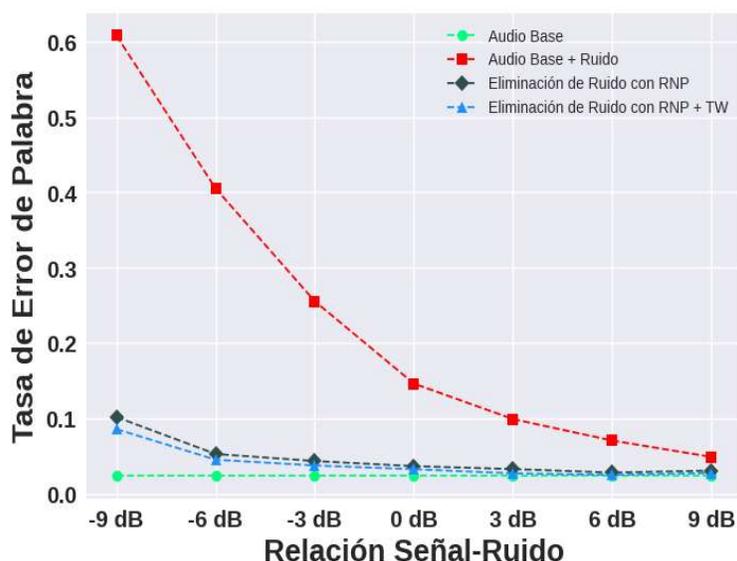


Figura 5.25 Comportamiento de la tasa de error de palabra - tipo de ruido: escritura de teclado

En la tabla 5.21 se presentan los resultados específicos de la tasa de error de palabra para cada nivel de SNR al tratar el ruido de escritura de teclado. Para los audios con este tipo de ruido, se obtuvo una tasa de error de palabra promedio de 0.2332. Al aplicar la red neuronal profunda, este error se reduce a 0.0464, y finalmente, al utilizar tanto la red neuronal profunda como la transformada Wavelet, se logra una disminución adicional de la tasa de error de palabra a 0.0399.

Tabla 5.21 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: escritura de teclado

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.6080	0.1016	0.0856
-6 dB	0.4051	0.0526	0.0450
-3 dB	0.2554	0.0432	0.0374
0 dB	0.1460	0.0364	0.0324
3 dB	0.0989	0.0326	0.0268
6 dB	0.0704	0.0279	0.0249
9 dB	0.0487	0.0306	0.0274
Promedio	0.2332	0.0464	0.0399

5.4.2 Resultados para datos objetivo

En esta sección se analiza la tasa de error de palabra en el dominio objetivo, centrándonos en los ruidos clasificados como no estacionarios. Se evaluará el impacto de estos ruidos en el reconocimiento automático de voz y se presentarán los resultados correspondientes.

5.4.2.1 Tipo de ruido no estacionario – llanto de bebé

A continuación, en la figura 5.26, se presenta el comportamiento de la tasa de error de palabra al tratar el ruido de llanto de bebé. Como se puede observar en el gráfico, la línea roja representa la tasa de error de palabra cuando el audio contiene este ruido, mientras que las líneas gris y azul representan la disminución efectiva de este ruido mediante las técnicas propuestas.

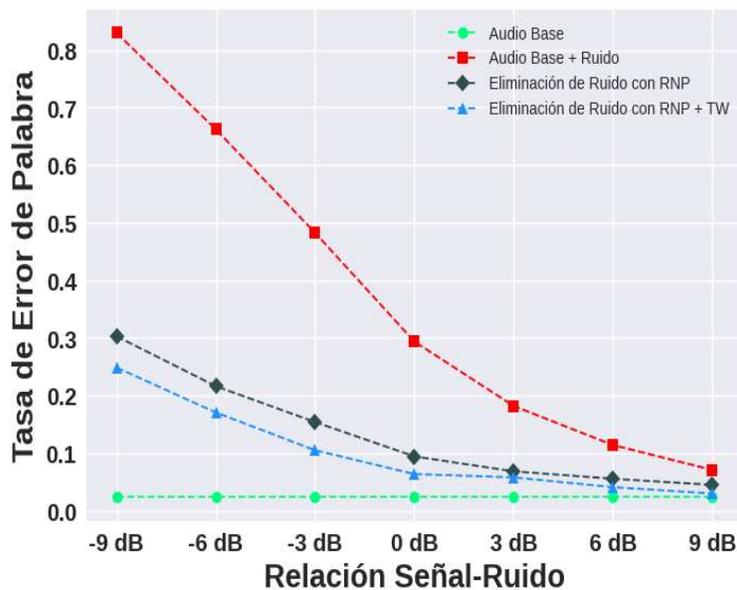


Figura 5.26 Comportamiento de la tasa de error de palabra - tipo de ruido: llanto de bebe

En la tabla 5.22 se presentan los resultados específicos de la tasa de error de palabra para cada nivel de SNR al tratar el ruido de llanto de bebé. Se observa que, para los audios con este tipo de ruido, se obtiene un error de palabra promedio de 0.3764. Sin embargo, al aplicar la red neuronal profunda, se logra reducir este error a 0.1336. Además, al combinar la red neuronal profunda con la transformada Wavelet, se alcanza una disminución adicional de la tasa de error de palabra, llegando a 0.1023.

5.4.2.2 Tipo de ruido no estacionario – fiesta con multitud de gente

En la figura 5.26 se presenta la tasa de error de palabra al mitigar el ruido de tipo no estacionario de multitud de gente. En el gráfico, se muestran cuatro líneas de diferentes colores. La línea verde representa la tasa de error de palabra cuando los audios no tienen ningún tipo de ruido, mientras que la línea roja representa la tasa de error de palabra cuando se introduce este tipo de ruido, que como se puede observar, resulta en un error significativamente mayor en

comparación con la línea base. Además, se muestra la mejora en el reconocimiento de voz mediante la aplicación de la red neuronal, representada por la línea gris, y la implementación conjunta de la red neuronal y la transformada Wavelet, representada por la línea azul. Estas técnicas han logrado una mejora ligera en comparación con los audios que contienen ruido en los diferentes niveles de SNR.

Tabla 5.22 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: llanto de bebe

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.8298	0.3028	0.2486
-6 dB	0.6622	0.2162	0.1706
-3 dB	0.4826	0.1539	0.1049
0 dB	0.2937	0.0943	0.0637
3 dB	0.1818	0.0684	0.0579
6 dB	0.1140	0.0554	0.0409
9 dB	0.0709	0.0448	0.0298
Promedio	0.3764	0.1336	0.1023

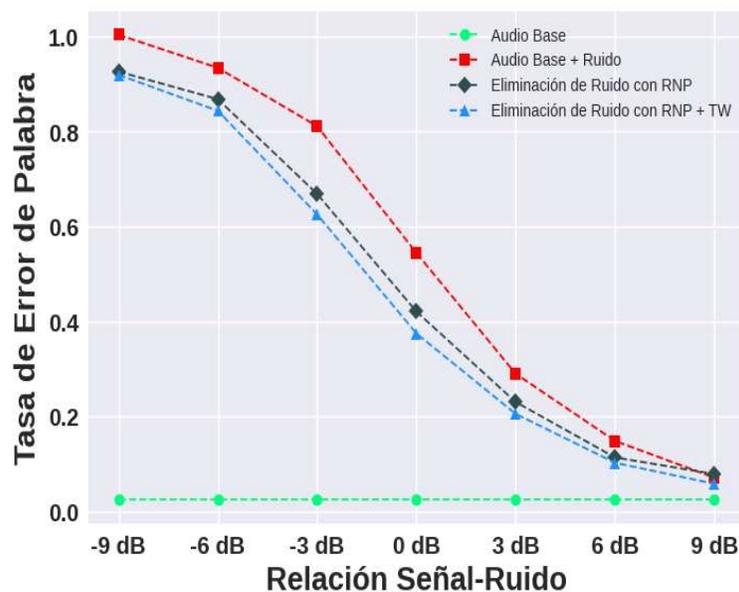


Figura 5.27 Comportamiento de la tasa de error de palabra - tipo de ruido: multitud de gente

En la tabla 5.23 se presentan los resultados de la tasa de error de palabra para la eliminación de ruido de tipo multitud de gente en diferentes niveles de SNR. En primer lugar, se muestra la tasa de error de palabra promedio para los audios que contienen este tipo de ruido, la cual es de 0.5428. Luego, al aplicar la red neuronal profunda, se observa una mejora moderada, reduciendo la tasa de error de palabra en promedio a 0.4718. Por último, al aplicar la transformada Wavelet en conjunto con la red neuronal profunda, se logra una leve mejora adicional, obteniendo una tasa de error de palabra de 0.4465.

Tabla 5.23 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: multitud de gente

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	1.0036	0.9260	0.9192
-6 dB	0.9340	0.8679	0.8435
-3 dB	0.8111	0.6682	0.6251
0 dB	0.5439	0.4204	0.3742
3 dB	0.2893	0.2297	0.2047
6 dB	0.1478	0.1129	0.1017
9 dB	0.0705	0.0780	0.0574
Promedio	0.5428	0.4718	0.4465

5.4.2.3 Tipo de ruido no estacionario – campanas de iglesia

En la figura 5.28 se muestra el comportamiento de la tasa de error de palabra al tratar el ruido de campanas de iglesia. La línea en color rojo representa el habla contaminada con este ruido, y se observa que la tasa de error de palabra es significativamente mayor que en los audios base. Sin embargo, al aplicar las técnicas de eliminación de ruido propuestas, se logra reducir considerablemente la tasa de error de palabra.

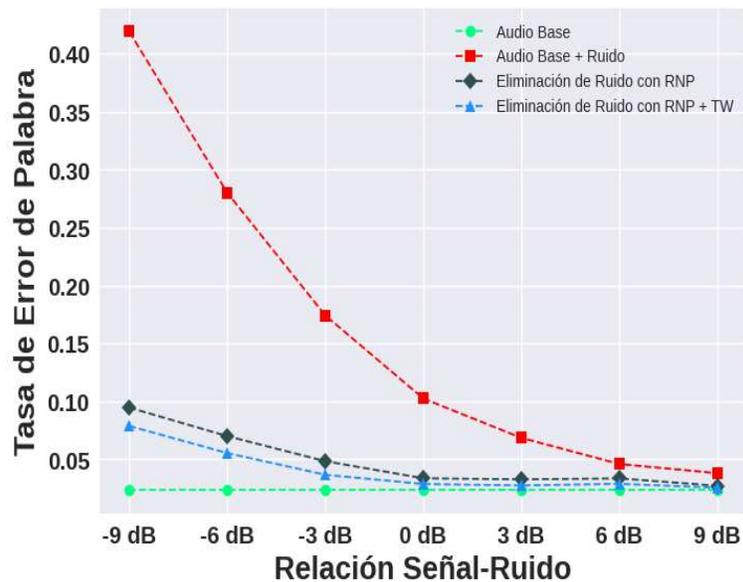


Figura 5.28 Comportamiento de la tasa de error de palabra - tipo de ruido: campanas de iglesia

En la tabla 5.24 se presentan los resultados numéricos de la tasa de error de palabra al tratar el ruido de campanas de iglesia en diferentes niveles de SNR. En general, se observa que este tipo de ruido tiene una tasa de error de palabra promedio de 0.1611. Sin embargo, al aplicar la red neuronal, se logra reducir la tasa de error de palabra a 0.0485. Además, al utilizar la transformada Wavelet como procesamiento adicional, se consigue disminuir aún más este error, alcanzando una tasa de error de palabra de 0.0401.

Tabla 5.24 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: campanas de iglesia

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.4191	0.0946	0.0788
-6 dB	0.2799	0.0700	0.0553
-3 dB	0.1742	0.0484	0.0366
0 dB	0.1024	0.0336	0.0287
3 dB	0.0685	0.0327	0.0274
6 dB	0.0459	0.0335	0.0288
9 dB	0.0380	0.0271	0.0255
Promedio	0.1611	0.0485	0.0401

5.4.2.4 Tipo de ruido no estacionario – mormullos en cafetería

En el gráfico de la figura 5.29 se muestra el comportamiento de la tasa de error de palabra en presencia del ruido de murmullos de cafetería. La línea roja representa la tasa de error de palabra cuando el habla está afectada por este tipo de ruido, mientras que la línea gris muestra la mejora obtenida al aplicar la red neuronal profunda para eliminar el ruido. Además, se muestra el efecto de aplicar la transformada Wavelet a este último proceso mencionado. El comportamiento se analiza en diferentes niveles de SNR, y se observa cómo las mejoras logran reducir ligeramente este tipo de ruido.

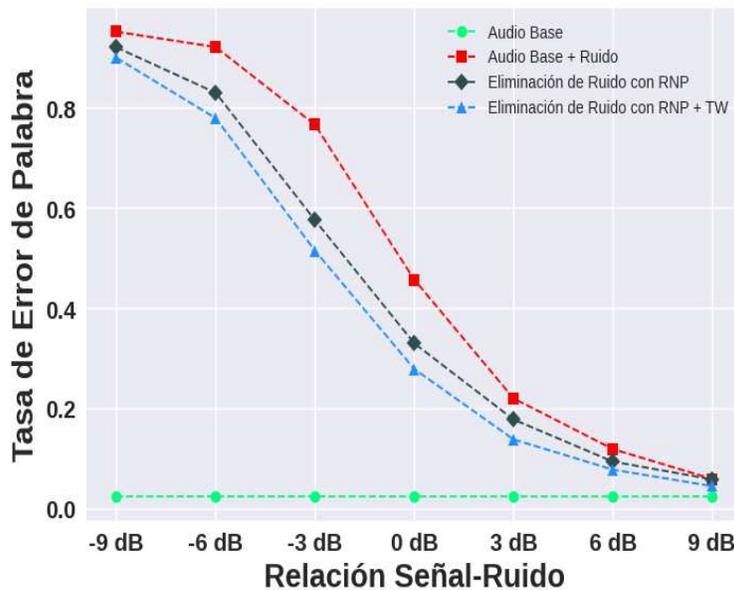


Figura 5.29 Comportamiento de la tasa de error de palabra - tipo de ruido: mormullos de cafetería

En la tabla 5.25 se presentan los resultados numéricos de la tasa de error de palabra al tratar el ruido de murmullos de cafetería. En promedio, se obtuvo una tasa de error de palabra de 0.4992 cuando los audios contenían este tipo de ruido. Sin embargo, al eliminar el ruido de la señal de voz, se logró reducir la tasa de error de palabra a 0.4266. Además, se observó una mejora adicional

al disminuir la tasa de error de palabra en 0.3905.

Tabla 5.25 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: mormullos de cafetería

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.9520	0.9208	0.8999
-6 dB	0.9215	0.8301	0.7794
-3 dB	0.7662	0.5757	0.5148
0 dB	0.4571	0.3299	0.2787
3 dB	0.2200	0.1778	0.1382
6 dB	0.1188	0.0938	0.0778
9 dB	0.0592	0.0583	0.0449
Promedio	0.4992	0.4266	0.3905

5.4.2.5 Tipo de ruido no estacionario – Helicóptero

En la figura 5.30 se presenta el comportamiento de la tasa de error de palabra en los diferentes niveles de SNR para el ruido de helicóptero. La línea roja representa el habla con ruido, mientras que las líneas en color gris y azul representan la disminución de este tipo de ruido. Se puede apreciar una mejora significativa en comparación con el habla ruidosa, lo que se traduce en una reducción de la tasa de error de palabra.

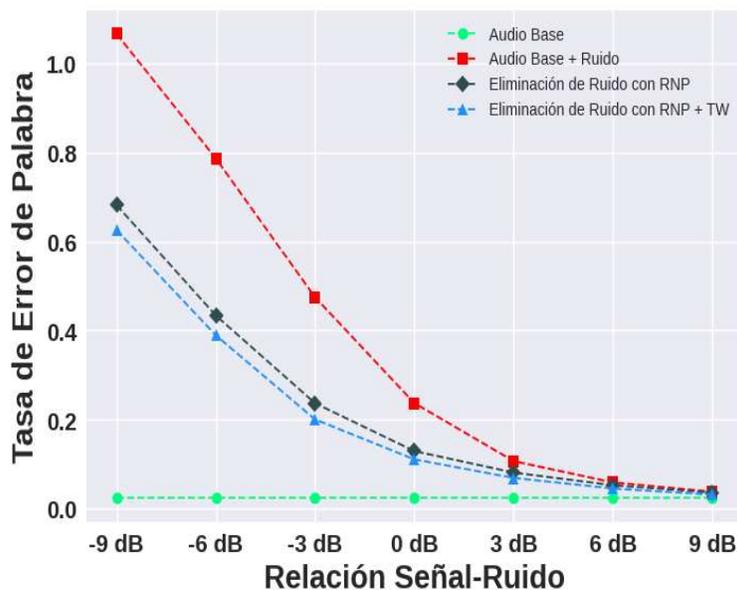


Figura 5.30 Comportamiento de la tasa de error de palabra - tipo de ruido: helicóptero

En la tabla 5.26 se presentan los resultados específicos obtenidos para cada nivel de SNR en el tratamiento del ruido de helicóptero. Se observa que el error promedio obtenido cuando el habla está contaminada con este ruido es de 0.3953. Sin embargo, al aplicar el procesamiento de la red neuronal, este error disminuye a 0.2358. Además, al incorporar el procesamiento adicional de la

transformada Wavelet, se logra una disminución adicional en la tasa de error de palabra, alcanzando un valor de 0.2102.

Tabla 5.26 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: helicóptero

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	1.0667	0.6819	0.6258
-6 dB	0.7867	0.4339	0.3894
-3 dB	0.4747	0.2367	0.2003
0 dB	0.2370	0.1292	0.1106
3 dB	0.1067	0.0810	0.0690
6 dB	0.058	0.0525	0.0452
9 dB	0.0379	0.0359	0.0315
Promedio	0.3953	0.2358	0.2102

5.4.2.6 Tipo de ruido no estacionario – personas hablando

En la figura 5.31 se muestra el comportamiento de la tasa de error de palabra en los diferentes niveles de SNR. La línea roja representa el habla con ruido, mientras que las líneas en color gris y azul representan la disminución de este tipo de ruido. Es evidente que se observa una mejora significativa en comparación con el habla ruidosa, lo que resulta en una disminución en la tasa de error de palabra.

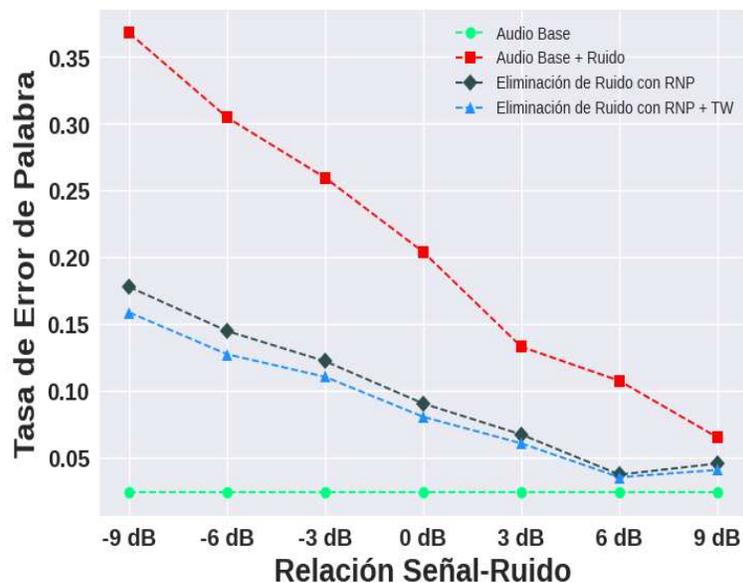


Figura 5.31 Comportamiento de la tasa de error de palabra - tipo de ruido: personas hablando

En la tabla 5.27 se muestran los resultados obtenidos para cada uno de los niveles de SNR en el tratamiento del ruido de personas hablando. El error promedio obtenido cuando el habla está contaminada con este ruido es de 0.2056. Sin embargo, al aplicar el procesamiento de la red neuronal, se logra reducir este error a 0.0974. Posteriormente, al aplicar la transformada Wavelet

como último procesamiento, se logra una disminución adicional en la tasa de error de palabra, alcanzando un valor de 0.0871.

Tabla 5.27 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: personas hablando

SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.3679	0.1774	0.1583
-6 dB	0.3045	0.1443	0.1269
-3 dB	0.2593	0.1217	0.1100
0 dB	0.2036	0.0900	0.0800
3 dB	0.1326	0.0668	0.0601
6 dB	0.1071	0.0367	0.0348
9 dB	0.0646	0.0450	0.0402
Promedio	0.2056	0.0974	0.0871

5.4.2.7 Tipo de ruido no estacionario – ladrido de perro

En la figura 5.32 se muestra el comportamiento de la tasa de error de palabra al eliminar el ruido de ladrido de perro. En esta ilustración, se destaca el error asociado cuando los audios del habla están contaminados con este tipo de ruido, representado por la línea en color rojo. Por otro lado, las líneas en color gris y azul representan las técnicas aplicadas para reducir este tipo de ruido. Dichas técnicas muestran una mejora significativa en la disminución del error, logrando una mejor calidad en el reconocimiento del habla.

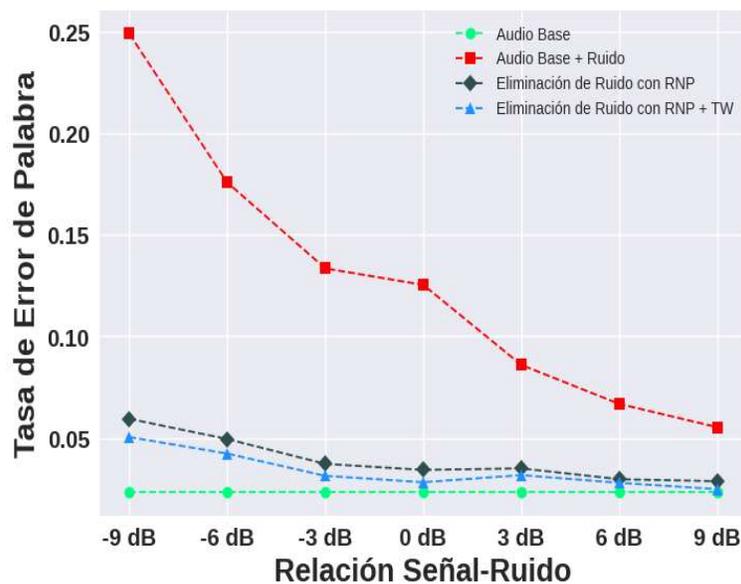


Figura 5.32 Comportamiento de la tasa de error de palabra - tipo de ruido: ladrido de perro

Los resultados numéricos de la evaluación del reconocedor automático de voz, en términos de la tasa de error de palabra, se presentan en la tabla 5.28 para el ruido de ladrido de perro en

diferentes niveles de SNR. En general, cuando el audio del habla está contaminado con este ruido, se obtiene un error promedio de 0.1272. Sin embargo, al aplicar la red neuronal profunda, se logra reducir este error a 0.0391. Además, al agregar el procesamiento adicional de la transformada Wavelet, se observa una ligera disminución adicional del error, llegando a 0.0338. Estos resultados demuestran la efectividad de las técnicas propuestas en la mejora del reconocimiento del habla en presencia de ruido de ladrido de perro.

Tabla 5.28 Tasa de error de palabra para cada nivel de SNR - tipo de ruido: ladridos de perro

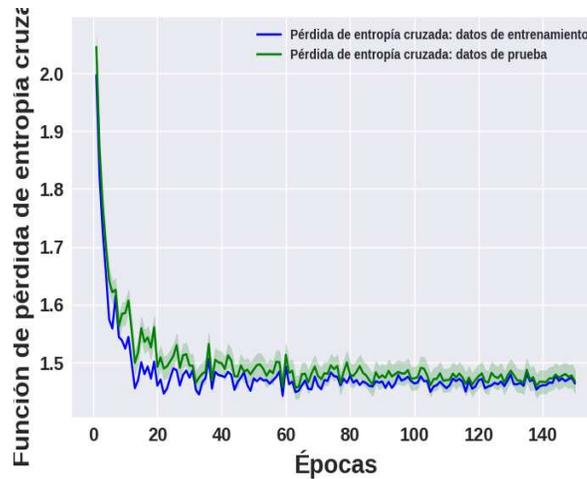
SNR	Audio ruidoso	Aplicando RNP	Aplicando RNP + TW
-9 dB	0.2488	0.0594	0.0504
-6 dB	0.1756	0.0495	0.0423
-3 dB	0.1334	0.0373	0.0314
0 dB	0.1253	0.0343	0.0282
3 dB	0.0859	0.0350	0.0317
6 dB	0.0667	0.0297	0.0279
9 dB	0.0553	0.0287	0.0248
Promedio	0.1272	0.0391	0.0338

5.5 Identificación de locutor

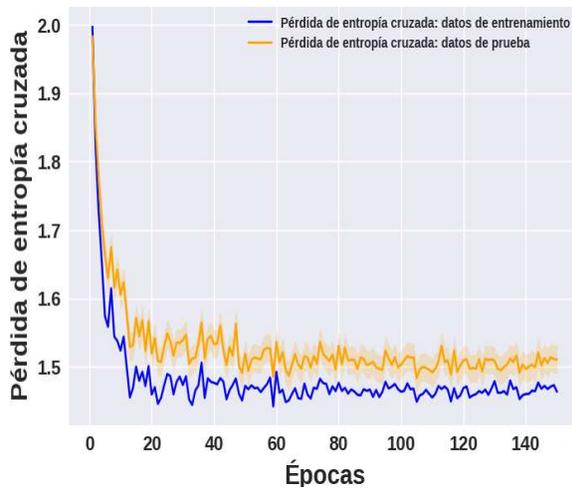
En la figura 5.33 se presenta el comportamiento de la función de pérdida de entropía cruzada durante cada época de entrenamiento de la red neuronal utilizada para la identificación de locutor. En este gráfico se muestran cuatro curvas distintas que corresponden a los siguientes conjuntos de datos: entrenamiento, prueba, datos prueba con eliminación de ruido y datos prueba con diferentes niveles de ruido en su relación señal-ruido. Es importante destacar las diferencias observadas al someter los audios de habla a diferentes condiciones. En primer lugar, en la figura 5.33 (a) se encuentra la curva correspondiente a los datos base, donde se puede apreciar un comportamiento de pérdida adecuado durante el entrenamiento de la red neuronal. Sin embargo, al considerar los datos con ruido, se observa en la figura 5.33 (c) una notable diferencia en la función de pérdida. La curva correspondiente a los datos con ruido muestra una mayor discrepancia y un aumento significativo en la pérdida, lo cual indica una dificultad para la red neuronal al tratar con el habla ruidosa. Por otro lado, en la figura 5.33 (b) los datos con eliminación de ruido presentan una curva de pérdida más favorable en comparación con los datos con ruido. Esto sugiere que la aplicación de la técnica de eliminación de ruido de la propuesta de trabajo ha permitido mejorar la capacidad de la red neuronal para procesar y clasificar los audios de cada locutor.

En la figura 5.34 se presenta el comportamiento de la métrica de exactitud durante el entrenamiento de la red neuronal profunda. Cada gráfico muestra la evolución de la exactitud a lo largo de las distintas épocas de entrenamiento, y la línea azul representa los datos de entrenamiento. En el primer caso, representado en la figura 5.34 (a), se evalúan los datos de prueba base. Aquí se observa que la exactitud alcanza un valor de 0.9900, lo que indica un alto nivel de precisión en la clasificación de los locutores. En el segundo caso, ilustrado en la figura 5.34 (b), se aplicó la eliminación de ruido a los audios. Se puede apreciar que la exactitud mejora, alcanzando un valor de 0.9950. Esto demuestra que la eliminación de ruido contribuye a una mejor clasificación de los locutores y aumenta la precisión del modelo. Sin embargo, en el tercer caso, mostrado en la figura 5.34 (c), se introdujo ruido en los audios. Aquí se observa una significativa

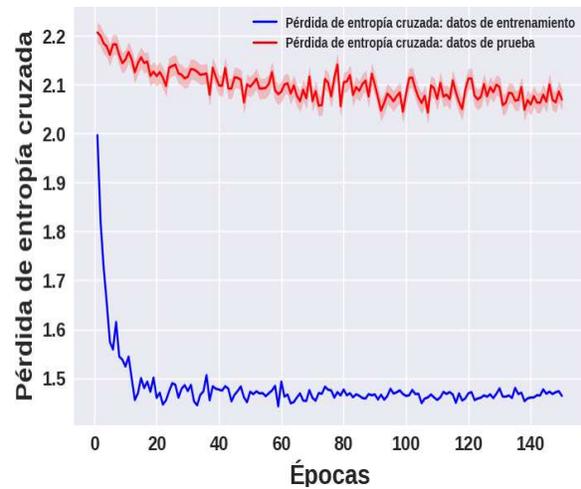
disminución en la exactitud, obteniendo un valor de 0.4350. Esto indica que el modelo es incapaz de reconocer adecuadamente a los locutores cuando se presentan condiciones de habla ruidosa, lo cual afecta negativamente la precisión de la clasificación.



a) Función de pérdida de entropía cruzada datos de entrenamiento base vs. datos de prueba base



b) Función de pérdida de entropía cruzada datos de entrenamiento base vs. datos de prueba con eliminación de ruido

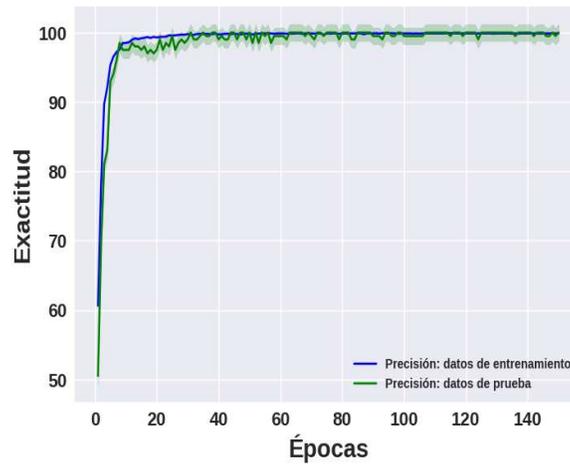


c) Función de pérdida de entropía cruzada datos de entrenamiento base vs. datos de prueba con ruido

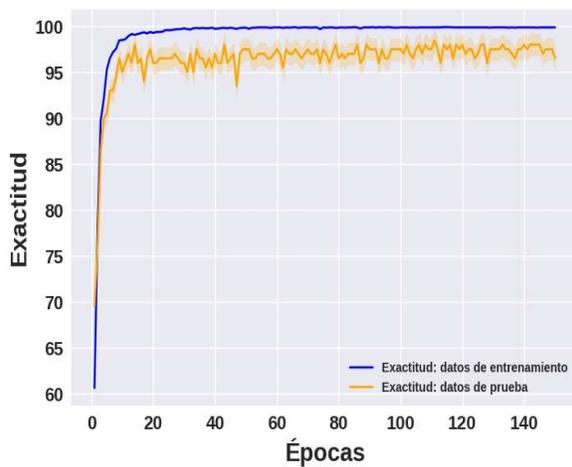
Figura 5.33 Comparación del comportamiento de la función de pérdida de entropía cruzada durante el entrenamiento de la red neuronal

En la tabla 5.29 se presentan las métricas utilizadas para evaluar el modelo en general de identificación de locutor. En primer lugar, se observa que con los audios de habla base, el modelo alcanza un 0.9900 de precisión, recall y F1-Score, mientras que la especificidad alcanza un valor de 0.9988. En segundo lugar, al aplicar la red neuronal para eliminación de ruido contenido en la señal de voz, se obtiene una mejora en las métricas con un valor de 0.9950 en precisión, recall y F1-Score, y una especificidad de 0.9990. Por otro lado, al tratar con audios con incrustación de ruido, se observa una degradación significativa en todas las métricas, alcanzando un valor de 0.4350 en precisión, recall y F1-Score, y una especificidad de 0.9372. Estos datos demuestran

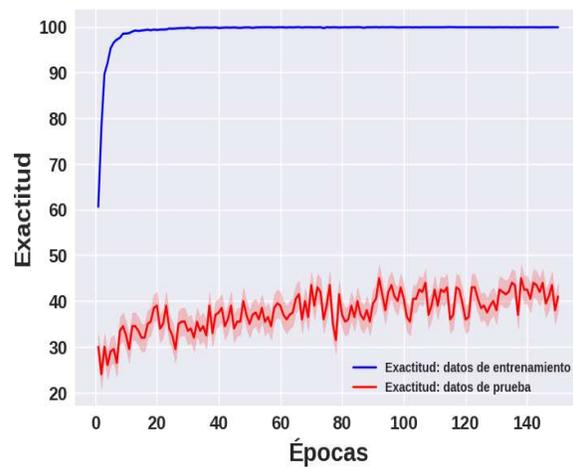
cómo los sistemas de identificación de locutor se ven afectados considerablemente por la presencia de ruido, lo cual dificulta su capacidad para realizar una clasificación precisa, ya que en este tipo de escenarios la precisión no supera el 50%.



a) Métrica de exactitud datos de entrenamiento base vs. datos de prueba base



b) Métrica de exactitud datos de entrenamiento base vs. datos de prueba con eliminación de ruido



c) Métrica de exactitud datos de entrenamiento base vs. datos de prueba con ruido

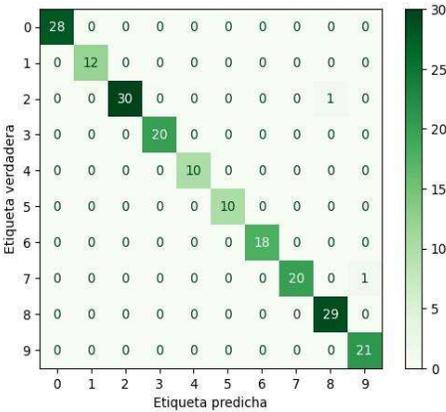
Figura 5.34 Comparación del comportamiento de la métrica de exactitud durante el entrenamiento de la red neuronal

Tabla 5.29 Métricas de evaluación para el modelo de identificación de locutor

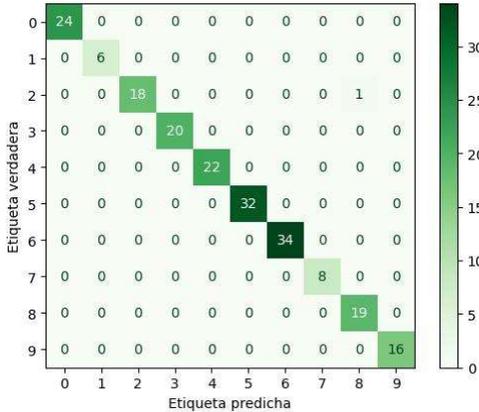
Métrica	Audio base	Audio con ruido	Audio con eliminación de ruido
Exactitud	0.9900	0.4350	0.9950
Recall	0.9900	0.4350	0.9950
Especificidad	0.9988	0.9372	0.9994
F1-Score	0.9900	0.4350	0.9950

En la figura 5.35 se presentan los resultados de las matrices de confusión correspondientes a

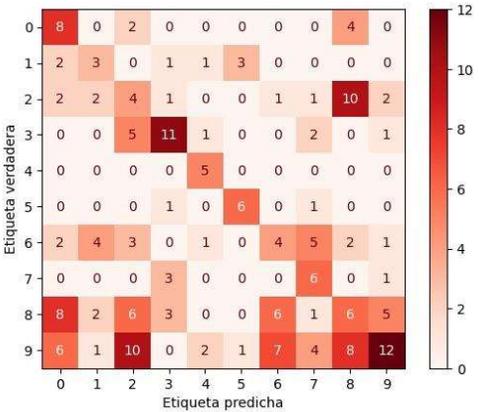
los distintos casos analizados: los audios base del corpus, los audios con eliminación de ruido y los audios del habla con ruido. Esta figura ofrece una visión del desempeño en los diferentes escenarios mencionados. En la figura 5.35 (a) se muestra la matriz de confusión de los audios base, donde se puede observar en la diagonal principal cómo los audios son clasificados correctamente para cada locutor. Esto indica que el modelo es capaz de identificar a los locutores de forma precisa en este caso. Por otro lado, en la figura 5.35 (b) se presentan los resultados de la matriz de confusión cuando se aplicó la red neuronal profunda junto a la transformada Wavelet para tratar los audios con eliminación de ruido. Aquí se observa un mayor acierto en los resultados, ya que la diagonal principal de la matriz de confusión indica que la mayoría de los locutores son clasificados de forma correcta. Esto demuestra que la propuesta de este trabajo para la eliminación de ruido mejora la capacidad de identificación de los locutores. Finalmente, en la figura 5.35 (c) se muestra la matriz de confusión correspondiente a los audios con diferentes tipos de ruido añadido en sus diferentes niveles de SNR, donde se evidencia una disminución en la capacidad de clasificación. La diagonal principal tiende a degradarse, lo que indica que la red neuronal para identificación de locutor no es capaz de identificar correctamente a cada uno de los locutores en estos escenarios con ruido. Esta métrica ofrece una visión general de los resultados obtenidos de la tabla 5.29.



a) Matriz de confusión con audios base de los locutores



b) Matriz de confusión con audios con eliminación de ruido de los locutores



c) Matriz de confusión con audios con ruido de los locutores

Figura 5.35 Matrices de confusión para cada conjunto de datos de prueba

Capítulo 6 Conclusiones

Este estudio se centró en la aplicación de técnicas de aprendizaje profundo y adaptación de dominio, junto con el filtrado basado en la transformada Wavelet, con el objetivo de mejorar la calidad y la inteligibilidad de la señal de voz en entornos ruidosos. Se llevó a cabo un análisis exhaustivo del efecto de distintos tipos de ruido, tanto estacionario como no estacionario, en la señal de voz, considerando diferentes niveles de relación señal-ruido (SNR) [-9,-6,-3,0,3,6,9] dB, basado en el estado del arte actual. Se observó que la mayoría de los trabajos existentes se centran en la eliminación de ruido a partir de un SNR de 0dB en adelante, mientras que este estudio logró mejoras iguales o superiores en niveles de SNR por debajo de 0dB en comparación con el estado del arte. Esto proporciona una visión específica de los tipos de ruido que se pueden mitigar en la señal de voz, obteniendo una mayor eliminación de ruido.

Cabe destacar que, en algunos casos, la combinación de la red neuronal profunda con adaptación de dominio y la aplicación de la transformada Wavelet, no logró mitigar suficientemente el ruido para mostrar mejoras en el sistema de reconocimiento automático de voz implementado. Un aspecto importante abordado en este estudio fue la adaptación de la red neuronal profunda y la técnica de filtrado basado en la transformada Wavelet para la eliminación de ruido. Inicialmente, cuando se aplicaba primero la transformada Wavelet, suprimía gran cantidad de información de la señal de voz, por lo que se optó por aplicar primero la red neuronal profunda y luego la técnica de filtrado de la transformada Wavelet. Además, se enfrentó el desafío de evitar la degradación de la señal de voz al aplicar un segundo proceso de filtrado, ya que esto podría llevar a la pérdida de información y a una disminución en la calidad general de la señal. La configuración adecuada del filtrado basado en la transformada Wavelet desempeñó un papel crucial en este estudio para lograr una mejora adicional en calidad de la señal. Durante las experimentaciones, la selección del tipo de onda, los niveles de descomposición y el tipo de filtro fueron aspectos primordiales, ya que algunas configuraciones demostraron una degradación en las métricas de STOI y PESQ de la señal de voz.

Mediante una cuidadosa aplicación de la técnica de filtrado basado en la transformada Wavelet, fue posible preservar la inteligibilidad obtenida por la red neuronal profunda y mejorar la calidad perceptual de la señal de habla. En términos de resultados, se logró aumentar las métricas de evaluación del habla, especialmente el STOI con un promedio del 20% y el PESQ con una mejora promedio del 9% en el corpus de datos. Estos resultados fueron obtenidos mediante la implementación de una red neuronal profunda capaz de eliminar el ruido en un dominio fuente y un dominio destino. Los beneficios de esta mejora en la señal de voz se reflejaron en una menor tasa de error de palabra del 14.24% en un sistema de RAV en comparación con el habla ruidosa. Además, se observó un aumento en las métricas de precisión, recall, especificidad y F1-score, con un promedio del 99% en la tarea de identificación de locutor en diversos entornos ruidosos.

6.1 Objetivos alcanzados

Durante el desarrollo de este trabajo, se lograron alcanzar los objetivos establecidos en la sección 1.6.

- El primer objetivo consistió en analizar en profundidad el efecto del ruido presente en la señal de voz en diferentes niveles de SNR. Para ello, se estudió el ruido añadido a la señal de voz en los niveles de SNR de [-9, -6, -3, 0, 3, 6, 9] dB, tal como se describió en el estado

del arte. Este enfoque fue crucial, ya que la mayoría de los trabajos existentes para mitigar el ruido se centran en niveles de SNR a partir de 0 dB. Sin embargo, se buscó abordar los niveles inferiores a 0 dB, ya que tienen una mayor influencia en la inteligibilidad del habla cuando se mezclan con la señal. Aunque algunos tipos de ruido son inherentemente difíciles de mitigar, como se observó tanto en este trabajo como en el estado del arte para casos como viento, cabina, murmullos de gente, fiestas con multitudes y ruido de vehículos; donde se lograron mejoras significativas en las métricas de STOI y PESQ. Sin embargo, no se obtuvo una disminución significativa en la tasa de error de palabra en el reconocimiento de voz.

- El segundo objetivo fue definir las técnicas de aprendizaje profundo y filtrado. Esto resultó fundamental, ya que se observó que, al someter la señal de voz a un segundo procesamiento, se podía perder información crucial. Para abordar esto, se diseñó una red neuronal profunda con adaptación de dominio capaz de eliminar diferentes tipos de ruido en la señal de voz utilizando el Transporte Óptimo. Una vez que se evaluaron los resultados y se reflejaron en las métricas esperadas, se propuso el uso de la transformada Wavelet para mejorar la calidad de la señal de voz. Durante la definición de este segundo procesamiento, se exploraron varias configuraciones, siendo la descrita en la sección 4.7 la que mejor se ajustó, ya que mantuvo en promedio el STOI y logró una mejora en el PESQ. Esta leve mejora se reflejó especialmente en el RAV, donde se logró reducir la tasa de error de palabra.
- El tercer objetivo consistió en evaluar las técnicas aplicadas en el reconocimiento del habla, y se obtuvieron mejoras significativas. En primer lugar, se logró incrementar el STOI y PESQ en un 20% y 9% respectivamente. Además, se observó una mejora del 14.24% en la tasa de error de palabra en el reconocimiento automático de voz en comparación con el habla ruidosa. Por último, se logró una mejora del 55% en la identificación de locutor en comparación con el habla ruidosa.
- Finalmente, se alcanzó el cuarto objetivo, ya que estas mejoras superaron el 5% establecido en la hipótesis planteada.

6.2 Hipótesis/proposiciones demostradas

La combinación de técnicas de filtrado y adaptación de dominio en redes neuronales profundas, utilizando el Transporte Óptimo, ha sido fundamental para mejorar el reconocimiento del habla en entornos ruidosos. La implementación de esta combinación ha demostrado mejoras significativas en las métricas, como el STOI y PESQ, alcanzando mejoras del 20% y 9%, respectivamente, en comparación con el habla ruidosa generada durante los experimentos. Además, se observó una mejora promedio del 14.24% en la tasa de error de palabra en comparación con el habla ruidosa. Este resultado destaca la efectividad de las técnicas aplicadas en la reducción de errores en el reconocimiento de palabras en entornos ruidosos. Adicionalmente, dado que el procesamiento para la eliminación de ruido puede llevar a la pérdida de información en la señal de voz, se aplicó la adaptación de dominio a la red neuronal usando el habla base y el habla obtenida del procesamiento aplicado para eliminación de ruido. Como resultado, se logró una mejora promedio del 55% en la identificación de locutor en comparación con el habla ruidosa. Por lo tanto, la hipótesis planteada no se rechaza.

En conjunto, estas mejoras representan un avance significativo en el reconocimiento del habla, con un aumento mayor al 5% propuesto respecto a la hipótesis planteada en las métricas

mencionadas. Estos resultados resaltan el potencial y la efectividad del enfoque propuesto en este estudio para abordar el reconocimiento del habla en condiciones de ruido, lo que abre nuevas posibilidades para investigaciones futuras en este campo.

6.3 Contribuciones de la investigación

El presente trabajo ofrece varias contribuciones significativas en el campo del reconocimiento del habla en entornos ruidosos. En primer lugar, se destaca por su enfoque diferencial al abordar tanto el reconocimiento automático de voz como la identificación de locutor en estos entornos. Además, implementamos métodos de aprendizaje profundo con adaptación de dominio utilizando el Transporte Óptimo en contextos de regresión y clasificación. Nuestro enfoque se basó en experimentaciones con el TO para alinear las etiquetas verdaderas con las predichas, respaldado por los hallazgos de Fatras et al. (2021), quienes describieron cómo los minimizadores de pérdida en modelos generativos convergen hacia el minimizador verdadero en cada lote. Implementamos un enfoque sencillo y escalable del TO, que se aplicó desde los lotes hasta los datos originales. Durante la fase de experimentación del TO, observamos un comportamiento interesante en cuanto a la convergencia de las redes neuronales en los contextos de regresión y clasificación.

Para el caso de regresión, utilizamos primero el error cuadrático medio como función de pérdida para mejorar el habla ruidosa, mientras que en la clasificación de locutores de texto independiente empleamos la entropía cruzada. Sin embargo, al reemplazar estas funciones con el plan de TO, según lo descrito por Fatras et al. (2021), obtuvimos una función de pérdida más robusta que tiene en cuenta los efectos del muestreo por lotes. Nuestra implementación del TO en el aprendizaje profundo demostró ser efectiva en la alineación de etiquetas y la mejora de la calidad del reconocimiento del habla en entornos ruidosos, además de aplicar una mejora adicional utilizando técnicas de filtrado. Estas observaciones abren nuevas perspectivas para futuras investigaciones en el campo.

6.4 Trabajos futuros

Como trabajo futuro, se puede considerar el siguiente enfoque para mejorar el reconocimiento del habla en entornos ruidosos y elevar la calidad y la inteligibilidad de la señal de voz. Basándonos en los resultados y experiencias de este trabajo, se propone explorar la combinación de la magnitud y la fase del espectro de Fourier, siguiendo los planteamientos de Choi et al. (2019) y Hu et al. (2020). Esta estrategia podría generar mejoras adicionales en el rendimiento del sistema. Además, se sugiere investigar otros métodos de eliminación de ruido, como aquellos basados en la restauración espectral, en lugar del enfoque de filtrado basado en la transformada Wavelet utilizado en este estudio. Estos métodos podrían ofrecer resultados prometedores y contribuir a reducir aún más la tasa de error de palabra, así como la precisión de identificación de locutor. Es importante destacar que las mejoras en el habla ruidosa continúan siendo relevantes en diversos ámbitos, como las interacciones con máquinas, la interpretación simultánea y la transcripción de audio, entre otros, como se ha mencionado por Chen et al. (2022). Por lo tanto, es esencial seguir investigando y desarrollando soluciones efectivas que aborden este desafío en la mejora de la comunicación de voz en condiciones ruidosas.

Referencias

- Abdullah, S., Zamani, M., & Demosthenous, A. (2021a). A Discrete Wavelet Transform-Based Voice Activity Detection and Noise Classification with Sub-Band Selection. *Proceedings - IEEE International Symposium on Circuits and Systems, 2021-May*. <https://doi.org/10.1109/ISCAS51556.2021.9401647>
- Abdullah, S., Zamani, M., & Demosthenous, A. (2021b). A Discrete Wavelet Transform-Based Voice Activity Detection and Noise Classification with Sub-Band Selection. *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5. <https://doi.org/10.1109/ISCAS51556.2021.9401647>
- Adeel, A., Gogate, M., & Hussain, A. (2020). Contextual Deep Learning-Based Audio-Visual Switching for Speech Enhancement in Real-World Environments. *Information Fusion*, 59, 163–170. <https://doi.org/10.1016/j.inffus.2019.08.008>
- Ahn, Y., Lee, S. J., & Shin, J. W. (2021). Cross-Corpus Speech Emotion Recognition Based on Few-Shot Learning and Domain Adaptation. *IEEE Signal Processing Letters*, 28, 1190–1194. <https://doi.org/10.1109/LSP.2021.3086395>
- Ali, M. H., Jaber, M. M., Abd, S. K., Rehman, A., Awan, M. J., Vitkutė-Adžgauskienė, D., Damaševičius, R., & Bahaj, S. A. (2022). Harris Hawks Sparse Auto-Encoder Networks for Automatic Speech Recognition System. *Applied Sciences (Switzerland)*, 12(3). <https://doi.org/10.3390/app12031091>
- Ambrosio, L., & Gigli, N. (2009). *A User's Guide to Optimal Transport*. <https://hal.archives-ouvertes.fr/hal-00769391>
- Anggun, W., Risanuri, H., & Agus, B. (2018). *Improvement of MFCC Feature Extraction Accuracy Using PCA in Indonesian Speech Recognition*.
- Ashar, A., Bhatti, M. S., & Mushtaq, U. (2020). Speaker Identification Using a Hybrid CNN-MFCC Approach. *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, 1–4. <https://doi.org/10.1109/ICETST49965.2020.9080730>
- Ashwin, J. S., & Manoharan, N. (2018). Audio Denoising Based on Short Time Fourier Transform. *Indonesian Journal of Electrical Engineering and Computer Science*, 9(1), 89–92. <https://doi.org/10.11591/ijeecs.v9.i1.pp89-92>
- Baffour, A. A., Qin, Z., Geng, J., Ding, Y., Deng, F., & Qin, Z. (2022). Generic Network for Domain Adaptation Based on Self-Supervised Learning and Deep Clustering. *Neurocomputing*, 476, 126–136. <https://doi.org/https://doi.org/10.1016/j.neucom.2021.12.099>
- Bansal, P., Kant, A., Kumar, S., Sharda, A., & Gupta, S. (2008). *Improved Hybrid Model of HMM/GMM for Speech Recognition*. <https://www.researchgate.net/publication/254427682>
- Becerra, A. (2017). *Reconocimiento de Voz a Través de Técnicas Híbridas Utilizando Modelos Markovianos y Nuevos Tipos de Redes Neuronales*.
- Bell, P., Fainberg, J., Klejch, O., Li, J., Renals, S., & Swietojanski, P. (2021). Adaptation Algorithms for Neural Network-Based Speech Recognition: An Overview. *IEEE Open Journal of Signal Processing*, 2, 33–66. <https://doi.org/10.1109/OJSP.2020.3045349>
- Bhat, G. S., Shankar, N., Reddy, C. K. A., & Panahi, I. M. S. (2019). A Real-Time Convolutional Neural Network Based Speech Enhancement for Hearing Impaired Listeners Using Smartphone. *IEEE Access*, 7, 78421–78433. <https://doi.org/10.1109/ACCESS.2019.2922370>
- Boyko, N., & Hrynyshyn, A. (2021). Using Recurrent Neural Network to Noise Absorption from Audio Files. *International Workshop on Computational & Information Technologies for Risk-Informed Systems*.
- Bunrit, S., Inkian, T., Kerdprasop, N., & Kerdprasop, K. (2019). Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network. *International Journal of Machine Learning and Computing*, 9, 143–148. <https://doi.org/10.18178/ijmlc.2019.9.2.778>
- Charniak, E. (1994). *Statistical Methods for Speech Recognition* (J. Goodman & H. C. Nusbaum, Eds.).

1994.

- Chehebar, G. N., & Groisman, P. (2021). *Transporté Óptimo y Baricentro con Distancia de Fermat*. Universidad de Buenos Aires.
- Chelali, F. zohra, Cherabit, Noureddine., Djeradi, Amar., & Falek, Leila. (2018). Wavelet Transform for Speech Compression and Denoising. *2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*, 1–7. <https://doi.org/10.1109/ICMCS.2018.8525996>
- Choi, H.-S., Kim, J.-H., Huh, J., Kim, A., Ha, J.-W., & Lee, K. (2019). *Phase-Aware Speech Enhancement with Deep Complex U-Net*.
- Chuang, F.-K., Wang, S.-S., Hung, J., Tsao, Y., & Fang, S.-H. (2019). *Speaker-Aware Deep Denoising Autoencoder with Embedded Speaker Identity for Speech Enhancement*. 3173–3177. <https://doi.org/10.21437/Interspeech.2019-2108>
- Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine*, 35(1), 53–65. <https://doi.org/10.1109/msp.2017.2765202>
- Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., & Courty, N. (2018). *DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation*.
- Daniel Jurafsky, J. H. Martin. (2009). *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall.
- Defossez, A., Synnaeve, G., & Adi, Y. (2020). *Real Time Speech Enhancement in the Waveform Domain*.
- Donahue, C., Li, B., & Prabhavalkar, R. (2018). *Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition*.
- Fan, C., Yi, J., Tao, J., Tian, Z., Liu, B., & Wen, Z. (2020). *Gated Recurrent Fusion with Joint Training Framework for Robust End-to-End Speech Recognition*. <http://arxiv.org/abs/2011.04249>
- Fatras, K., Cuturi, M., Christian, C., Sebag, M., Chapel, L., Courty, N., & Flamary, R. (2021). *Optimal Transport and Deep Learning: Learning from one another*. L'Université Bretagne Sud.
- Geron, A. (2017). *Hands-on Machine Learning with Scikit-Learn and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative Adversarial Networks. *Commun. ACM*, 63(11), 139–144. <https://doi.org/10.1145/3422622>
- Graves, A., Jaitly, N., & Mohamed, A. (2013). Hybrid Speech Recognition with Deep Bidirectional LSTM. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 273–278. <https://doi.org/10.1109/ASRU.2013.6707742>
- Graves, A., & Schmidhuber, J. (2005). Framewise Phoneme Classification with Bidirectional LSTM Networks. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, 4, 2047–2052 vol. 4. <https://doi.org/10.1109/IJCNN.2005.1556215>
- Haque, A. (2020). EC-GAN: Low-Sample Classification using Semi-Supervised Algorithms and GANs. *ArXiv, abs/2012.15864*.
- Harchaoui, W. (2020). *Learning Representations using Neural Networks and Optimal Transport*. <https://tel.archives-ouvertes.fr/tel-03370529>
- Haridas, A. V., Marimuthu, R., & Sivakumar, V. G. (2018). A Critical Review and Analysis on Techniques of Speech Recognition: The Road Ahead. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 22(1), 39–57. <https://doi.org/10.3233/KES-180374>
- Haton, J.-P. (1994). Problems and Solutions for Noisy Speech Recognition. *Journal de Physique IV Proceedings, 111(C5)*, 4. <https://doi.org/10.1051/jp4:1994592i>
- Hernández, R., Méndez, S., Mendoza, C., & Cuevas, R. (2017). *Fundamentos de Investigación* (1st ed.). 2017.
- Hidayat, R., & Winursito, A. (2020). A Modified MFCC for Improved Wavelet-Based Denoising on Robust Speech Recognition. *International Journal of Intelligent Engineering and Systems*, 14(1), 12–21. <https://doi.org/10.22266/IJIES2021.0228.02>
- Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., & Xie, L. (2020). *DCCRN: Deep*

- Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement.*
- Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR*, *abs/1508.01991*. <http://arxiv.org/abs/1508.01991>
- Hwang, D., Misra, A., Huo, Z., Siddhartha, N., Garg, S., Qiu, D., Sim, K. C., Strohman, T., Beaufays, F., & He, Y. (2022). *Large-Scale ASR Domain Adaptation using Self-and Semi-Supervised Learning.*
- Jinyu, L., Li, D., Yifan, G., & Reinhold, H.-U. (2014). An Overview of Noise-Robust Automatic Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4), 745–777.
- Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.
- Kehtarnavaz, N. (2008). Chapter 7 - Frequency Domain Processing. In N. Kehtarnavaz (Ed.), *Digital Signal Processing System Design (Second Edition)* (Second Edition, pp. 175–196). Academic Press. <https://doi.org/https://doi.org/10.1016/B978-0-12-374490-6.00007-6>
- Khdier, H., Jasim, W., & Aliesawi, S. (2021). Deep Learning Algorithms based Voiceprint Recognition System in Noisy Environment. *Journal of Physics: Conference Series*, 1804, 12042. <https://doi.org/10.1088/1742-6596/1804/1/012042>
- Khurana, S., Moritz, N., Hori, T., & Roux, J. Le. (2020). *Unsupervised Domain Adaptation for Speech Recognition via Uncertainty Driven Self-Training.* <http://arxiv.org/abs/2011.13439>
- Leglaive, S., Girin, L., & Horaud, R. (2019). Semi-supervised Multichannel Speech Enhancement with Variational Autoencoders and Non-negative Matrix Factorization. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 101–105. <https://doi.org/10.1109/ICASSP.2019.8683704>
- Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An Overview of Noise-Robust Automatic Speech Recognition. In *IEEE Transactions on Audio, Speech and Language Processing* (Vol. 22, Issue 4, pp. 745–777). Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/TASLP.2014.2304637>
- Li, Q., Yang, Y., Lan, T., Zhu, H., Wei, Q., Qiao, F., Liu, X., & Yang, H. (2020). MSP-MFCC: Energy-Efficient MFCC Feature Extraction Method with Mixed-Signal Processing Architecture for Wearable Speech Recognition Applications. *IEEE Access*, 8, 48720–48730. <https://doi.org/10.1109/ACCESS.2020.2979799>
- Li, Q., Zhu, H., Qiao, F., Wei, Q., Liu, X., & Yang, H. (2018). *Energy-efficient MFCC extraction architecture in mixed-signal domain for automatic speech recognition.* 138–140. <https://doi.org/10.1145/3232195.3232219>
- Li, Y., Sun, Y., Horoshenkov, K., & Naqvi, S. M. (2022). Domain Adaptation and Autoencoder-Based Unsupervised Speech Enhancement. *IEEE Transactions on Artificial Intelligence*, 3(1), 43–52. <https://doi.org/10.1109/tai.2021.3119927>
- Liao, C.-F., Tsao, Y., Lee, H.-Y., & Wang, H.-M. (2019). *Noise Adaptive Speech Enhancement using Domain Adversarial Training.*
- Lin, H.-Y., Tseng, H.-H., Lu, X., & Tsao, Y. (2021). *Unsupervised Noise Adaptive Speech Enhancement by Discriminator-Constrained Optimal Transport.* <http://arxiv.org/abs/2111.06316>
- Liu, B., Nie, S., Liang, S., Liu, W., Yu, M., Chen, L., Peng, S., & Li, C. (2019). Jointly adversarial enhancement training for robust end-to-end speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2019-September*, 491–495. <https://doi.org/10.21437/Interspeech.2019-1242>
- Meriane Brahim. (2021). Denoising and Enhancement Speech Signal Using Wavelet. *Journal of Information Systems and Telecommunication (JIST)*, 9(1). <https://doi.org/10.52547/jist.9.33.37>
- Mohamed, H., Hassan, S., Ouissam, Z., & Khalid, S. (2020). Amazigh digits through interactive speech recognition system in noisy environment. *International Journal of Speech Technology*, 23(1), 101–109.
- Multisensor. (2014). *Basic techniques for speech recognition, text analysis and concept detection.*
- Naing, H. M., Hidayat, R., Hartanto, R., & Miyayaga, Y. (2020). Discrete Wavelet Denoising into MFCC for Noise Suppressive in Automatic Speech Recognition System. *International Journal of Intelligent*

- Engineering and Systems*, 13, 74–82. <https://doi.org/10.22266/ijies2020.0430.08>
- Nossier, S. A., Wall, J., Moniri, M., Glackin, C., & Cannings, N. (2021). An Experimental Analysis of Deep Learning Architectures for Supervised Speech Enhancement. *Electronics*, 10(1). <https://doi.org/10.3390/electronics10010017>
- Nyein Thu, L., Win, A., & Ne Oo, H. (2008). IRJET-A Review for Reduction of Noise by Wavelet Transform in Audio Signals A Review for Reduction of Noise by Wavelet Transform in Audio Signals. *International Research Journal of Engineering and Technology*, 8128. www.irjet.net
- O’Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. *ArXiv*, [abs/1511.08458](https://arxiv.org/abs/1511.08458).
- Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR Corpus Based on Public Domain Audio Books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. <https://doi.org/10.1109/ICASSP.2015.7178964>
- Park, G., Cho, W., Kim, K.-S., & Lee, S. (2020). Speech Enhancement for Hearing Aids with Deep Learning on Environmental Noises. *Applied Sciences*, 10(17). <https://doi.org/10.3390/app10176077>
- Parveen, N., & Abdullah, K. (2019). Diversity Technique Using Discrete Wavelet Transform In OFDM System Internet of Things View project Design and Development of Non-uniform Planar Array View project. In *International Journal of Engineering and Advanced Technology (IJEAT)* (Issue 8). <https://www.researchgate.net/publication/332878743>
- Patil, R. (2015). Noise Reduction using Wavelet Transform and Singular Vector Decomposition. *Procedia Computer Science*, 54, 849–853. <https://doi.org/10.1016/j.procs.2015.06.099>
- Peyré, G. (2018). *Numerical Optimal Transport and its Applications*.
- Phan, H., McLoughlin, I. V., Pham, L., Chen, O. Y., Koch, P., Vos, M. De, & Mertins, A. (2020). Improving GANs for Speech Enhancement. *IEEE Signal Processing Letters*, 27, 1700–1704. <https://doi.org/10.1109/lsp.2020.3025020>
- Piczak, K. J. (2015). ESC: Dataset for Environmental Sound Classification. *Proceedings of the 23rd ACM International Conference on Multimedia*, 1015–1018. <https://doi.org/10.1145/2733373.2806390>
- Qin, X., Cai, D., & Li, M. (2019). Far-Field End-to-End Text-Dependent Speaker Verification Based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation. *Proc. Interspeech 2019*, 4045–4049. <https://doi.org/10.21437/Interspeech.2019-1542>
- Rezende, E. R. S., Ruppert, G. C. S., de Carvalho, T. J., Theóphilo, A., Ramos, F. T., & de Geus, P. L. (2018). *Malicious Software Classification Using VGG16 Deep Neural Network’s Bottleneck Features*.
- Risanuri, H., Agus, B., Sujoko, S., & Anggun, W. (2018). *Denoising Speech for MFCC Feature Extraction Using Wavelet Transformation in Speech Recognition*.
- Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual Evaluation of Speech Quality (PESQ): A New Method for Speech Quality Assessment of Telephone Networks and Codecs. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, 2, 749–752 vol.2. <https://doi.org/10.1109/ICASSP.2001.941023>
- Rosenberg, A. E. (1976). Automatic Speaker Verification: A Review. *Proceedings of the IEEE*, 64(4), 475–487. <https://doi.org/10.1109/PROC.1976.10156>
- Roy, S. K., Nicolson, A., & Paliwal, K. K. (2021). DeepLPC: A Deep Learning Approach to Augmented Kalman Filter-Based Single-Channel Speech Enhancement. *IEEE Access*, 9, 64524–64538. <https://doi.org/10.1109/ACCESS.2021.3075209>
- Salamon, J., Jacoby, C., & Bello, J. P. (2014). A Dataset and Taxonomy for Urban Sound Research. *Proceedings of the 22nd ACM International Conference on Multimedia*, 1041–1044. <https://doi.org/10.1145/2647868.2655045>
- Santambrogio, F. (2018). *A Short Story on Optimal Transport and its Many Applications*.
- Schrieber, J. (2019). *Algorithms for Optimal Transport and Wasserstein Distances*. Georg-August University School of Science.
- Sim, K. C., Narayanan, A., Misra, A., Tripathi, A., Pundak, G., Sainath, T. N., Haghani, P., Li, B., & Bacchiani, M. (2018). Domain adaptation using factorized hidden layer for robust automatic speech

- recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2018-September*, 892–896. <https://doi.org/10.21437/Interspeech.2018-2246>
- Singh, V., & Singh, S. (2015). *Audio Noise Reduction using Discrete Wavelet Transformation*. www.ijert.org
- Soe Naing, H. M., Hidayat, R., Hartanto, R., & Miyanaga, Y. (2020). Discrete Wavelet Denoising into MFCC for Noise Suppressive in Automatic Speech Recognition System. *International Journal of Intelligent Engineering and Systems*, 13(2), 74–82. <https://doi.org/10.22266/ijies2020.0430.08>
- Sokolov, A., & Savchenko, A. V. (2021). Gender Domain Adaptation for Automatic Speech Recognition. *SAMI 2021 - IEEE 19th World Symposium on Applied Machine Intelligence and Informatics, Proceedings*, 413–417. <https://doi.org/10.1109/SAMI50585.2021.9378626>
- Sun, S., Guo, P., Xie, L., & Hwang, M. Y. (2019). Adversarial Regularization for Attention Based End-to-End Robust Speech Recognition. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 27(11), 1826–1838. <https://doi.org/10.1109/TASLP.2019.2933146>
- Sun, S., Zhang, B., Xie, L., & Zhang, Y. (2017). An Unsupervised Deep Domain Adaptation Approach for Robust Speech Recognition. *Neurocomputing*, 257, 79–87. <https://doi.org/10.1016/j.neucom.2016.11.063>
- Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4214–4217. <https://doi.org/10.1109/ICASSP.2010.5495701>
- Tan, K., Zhang, X., & Wang, D. (2019). Real-time Speech Enhancement Using an Efficient Convolutional Recurrent Network for Dual-microphone Mobile Phones in Close-talk Scenarios. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5751–5755. <https://doi.org/10.1109/ICASSP.2019.8683385>
- Theckedath, D., & Sedamkar, R. (2020). Detecting Affect States Using VGG16, ResNet50 and SE-ResNet50 Networks. *SN Computer Science*, 1. <https://doi.org/10.1007/s42979-020-0114-9>
- Thiemann, J., Ito, N., & Vincent, E. (2013). *DEMAND: A Collection of Multi-Channel Recordings of Acoustic Noise in Diverse Environments*. Zenodo. <https://doi.org/10.5281/zenodo.1227121>
- Thu, L., Win, A., & Ne Oo, H. (2019). *A Review for Reduction of Noise by Wavelet Transform in Audio Signals*.
- Tsao, Y., Li, B., & Chai Sim, K. (2017). *An Investigation of Spectral Restoration Algorithms for Deep Neural Networks Based Noise Robust Speech Recognition*. <https://www.researchgate.net/publication/289255711>
- Vazhenina, D., & Markov, K. (2020). End-to-End Noisy Speech Recognition Using Fourier and Hilbert Spectrum Features. *Electronics*.
- Vikramjit, M., Hosung, N., Espy-Wilson, C. Y., & Saltzman, E. (2011). Articulatory Information for Noise Robust Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 1913–1924.
- Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., & Wang, F.-Y. (2017). Generative Adversarial Networks: Introduction and Outlook. *IEEE/CAA Journal of Automatica Sinica*, 4(4), 588–598. <https://doi.org/10.1109/JAS.2017.7510583>
- Wang, S., & Li, G. (2019). Overview of End-to-End Speech Recognition. *Journal of Physics: Conference Series*, 1187(5). <https://doi.org/10.1088/1742-6596/1187/5/052068>
- Wang, Z. Q., & Wang, D. L. (2016). A Joint Training Framework for Robust Automatic Speech Recognition. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 24(4), 796–806. <https://doi.org/10.1109/TASLP.2016.2528171>
- Watanabe, S. (2021). *ESPnet2 pretrained model, Shinji Watanabe/gigaspee_ch_asr_train_asr_raw_en_bpe5000_valid.acc.ave, fs=16k, lang=en*. Zenodo. <https://doi.org/10.5281/zenodo.4630406>
- Wei-Ning, H., Yu, Z., & James, G. (2017). *Unsupervised Domain Adaptation for Robust Speech*

- Recognition Via Variational Autoencoder-Base Data Augmentation.*
- Willett, D., & Rigou, G. (1997). *Hybrid NNHMM-Based Speech Recognition with a Discriminant Neural Feature Extraction.*
- Wu, J., & He, J. (2022). Domain Adaptation with Dynamic Open-Set Targets. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2039–2049. <https://doi.org/10.1145/3534678.3539235>
- Ye, F., & Yang, J. (2021). A Deep Neural Network Model for Speaker Identification. *Applied Sciences*, 11(8). <https://doi.org/10.3390/app11083603>
- Yu, D., & Deng, L. (2014). *Signals and Communication Technology Automatic Speech Recognition A Deep Learning Approach.* <http://www.springer.com/series/4748>
- Zhang, D. (2019). Wavelet Transform. In *Fundamentals of Image Data Mining: Analysis, Features, Classification and Retrieval* (pp. 35–44). Springer International Publishing. https://doi.org/10.1007/978-3-030-17989-2_3
- Zhao, F., Li, H., & Zhang, X. (2019). A Robust Text-independent Speaker Verification Method Based on Speech Separation and Deep Speaker. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6101–6105. <https://doi.org/10.1109/ICASSP.2019.8683762>
- Zhao, H., Zhang, S., Wu, G., Costeira, J. P., Moura, J. M. F., & Gordon, G. J. (2017). *Multiple Source Domain Adaptation with Adversarial Training of Neural Networks.*
- Zhao, X., Wang, Y., & Wang, D. (2014). Robust Speaker Identification in Noisy and Reverberant Conditions. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3997–4001. <https://doi.org/10.1109/ICASSP.2014.6854352>

Anexos

Se presento en el 1er Congreso Internacional de Mecatrónica, Control e Inteligencia Artificial (CIMCIA) un artículo que llevo por nombre: “Gender Classification and Speaker Identification Using Machine Learning Algorithms”, que se llevó a cabo del 9 al 11 de noviembre de 2022.



UNAM
UNAM
UNAM

Universidad Nacional Autónoma de México
Facultad de Estudios Superiores Cuautitlán
Departamento de Ingeniería

CIM
CIA

Otorgan la presente

Constancia

A: Emmanuel de J. Velásquez-Martínez, Aldonso Becerra-Sánchez, José I. De La Rosa-Vargas, Efrén González-Ramírez, Gustavo Zepeda-Valles, Armando Rodarte-Rodríguez, Nivia I. Escalante-García y J. Ernesto Olvera-González

Por su participación como ponente de la conferencia

Gender classification and speaker identification using machine learning algorithms

En el **1er. Congreso Internacional de Mecatrónica, Control e Inteligencia Artificial (CIMCIA)**,
realizado del 9 al 11 de noviembre de 2022.

“POR MI RAZA HABLARÁ EL ESPÍRITU”
Cuautitlán Izcalli, Estado de México, noviembre de 2022.



Folios: 202200434

Dr. David Quintanar Guerrero
Director

Se acepto el artículo titulado “Combinig Deep Learning with Domain Adaptation and Filtering Techniques for Speech Recognition in Noisy Environments” en IEEE Autumn Meeting on Power, Electronics and Computing en su edición ROPEC 2023.

