Universidad Autónoma de San Luis Potosí

Facultad de Ciencias

Detección Automática de Ciclos Respiratorios con Sibilancias Empleando Representaciones Tiempo-Frecuencia Clásicas y de Alta Resolución

Tesis para obtener el grado de

Maestría en Ciencias de la Vida

con Orientación en Bioingeniería

Presenta

Ing. Amaranta Perla Córdova Martínez

Asesores:

Dr. Bersaín Alexander Reyes

Facultad de Ciencias, Universidad Autónoma de San Luis Potosí, UASLP

Dr. Tomás Aljama Corrales

Universidad Autónoma Metropolitana, Unidad Iztapalapa, UAM-I

San Luis Potosí, S.L.P., febrero de 2024

A mis padres Perla y Julián

que, con su apoyo y amor incondicional,

me han guiado hasta el día de hoy.

Agradecimientos

Gracias a mi familia, mis padres Perla y Julián y mis hermanos Alejandra y Arturo por siempre guiarme, apoyarme y por todo su amor.

También quiero agradecer de todo corazón a mis asesores, el Dr. Bersaín, por toda su sabiduría, guía y comprensión durante este proceso, y al Dr. Tomás, por toda la paciencia y conocimientos.

Gracias a Eric, por darme todo su apoyo y cariño siempre.

Gracias a CONAHCYT por ser beneficiaria de una de sus becas (CVU/Becario: 842473/ 726320).

Detección Automática de Ciclos Respiratorios con Sibilancias Empleando Representaciones Tiempo-Frecuencia de Alta Resolución

Amaranta Perla Córdova Martínez
Universidad Autónoma de San Luis Potosí

RESUMEN

Los sonidos respiratorios (SR) digitalizados proveen información importante, para aplicaciones como la detección automática de sonidos, que puede auxiliar en el diagnóstico de patologías. Uno de los SR adventicios continuos comunes es la sibilancia, la cual está correlacionada con enfermedades como asma y enfermedad pulmonar obstructiva crónica (EPOC). En las últimas décadas, se han realizado esfuerzos para detectar automáticamente las sibilancias en los SR, e.g., mediante análisis de representaciones tiempo-frecuencia (*time-frequency representations*, TFR) clásicas como el espectrograma, el cual permite la visualización de cambios instantáneos en la frecuencia a lo largo del tiempo. Sin embargo, debido a sus características y a las limitaciones de las TFRs empleadas anteriormente, la detección de sibilancias aún presenta áreas de oportunidad. En este trabajo, proponemos realizar la detección de SR normales y SR que contienen sibilancias empleando cuatro distintas TFR: 1) Espectrograma (*Spectrogram*, SP), 2) Espectro de Hilbert-Huang (Hilbert-Huang *spectrum*, HHS), 3) Transformada Synchrosqueezing (*Synchrosqueezing transform*, SST) y 4) Distribución Wigner-Ville (Wigner-Ville *distribution*, WVD), así como una red neuronal convolucional pre-entrenada, *GoogleNet*, mediante transferencia de aprendizaje. El método propuesto fue evaluado en una base de datos pública denominada ICBHI *Challenge* 2017, la cual está compuesta por 920 sonidos respiratorios adquiridos de 126 participantes. La base de datos contiene 6898 ciclos respiratorios que incluyen sibilancias, crepitancias o una combinación de ambos, además de ciclos sin sonidos adventicios. En esta tesis, el método propuesto fue evaluado empleando 575 ciclos respiratorios que contienen sibilancias, pertenecientes a 45 pacientes, y 575 ciclos respiratorios normales, de 44 sujetos. El mejor resultado experimental alcanzó una precisión de 0.880, exactitud de 0.872 y sensibilidad de 0.860 para el conjunto de validación empleando el método SP. Así como una precisión de 0.666, exactitud de 0.705 y sensibilidad de 0.823 para el conjunto de prueba al aplicar el método SST. Al evaluar el conjunto de validación, el HHS obtuvo 0.779 de exactitud, 0.790 de sensibilidad, 0.767 de especificidad y 0.772 de precisión. Mientras que la WVD alcanzó 0.830 de exactitud, 0.833 de sensibilidad, 0.827 de especificidad y 0.828 de precisión con el mismo conjunto. Al comparar este desempeño con los reportados en la literatura, se encontró que los resultados obtenidos son por lo menos comparables con aquellos indicados en la mayoría de los trabajos del estado del arte. Consideramos que los resultados de esta tesis son prometedores en el área de detección de ciclos respiratorios con sibilancias en comparación con los procedimientos actuales y parecen apuntar a que las TFR de alta resolución, en especial la SST, es un buen contendiente contra el método tradicional, SP, y su implementación en un sistema posterior puede apoyar al diagnóstico de enfermedades respiratorias obstructivas.

Automatic Detection of Wheezing Respiratory Cycles using High-Resolution Time-Frequency Representations

Amaranta Perla Córdova Martínez
Universidad Autónoma de San Luis Potosí

ABSTRACT

Digitized respiratory sounds (RS) provide important information for applications such as automatic sound detection, which can assist in the diagnosis of pathologies. One of the common continuous adventitious RS is wheezing, which is correlated with diseases such as asthma and chronic obstructive pulmonary disease (COPD). In recent decades, efforts have been made to automatically detect wheeze in RS, e.g., through analysis of classical time-frequency representations (TFR) such as the spectrogram, which allows the visualization of instantaneous changes in frequency over time. However, due to its characteristics and the limitations of the TFR previously used, wheeze detection still presents areas of opportunity. In this work, we propose to perform the detection of normal RS and RS containing wheeze using four different TFRs: 1) Spectrogram (SP), 2) Hilbert-Huang Spectrum (HHS), 3) Synchrosqueezing Transform (SST) and 4) Wigner-Ville Distribution (WVD), as well as a pre-trained convolutional neural network, GoogleNet, through transfer learning. The proposed method was evaluated on a public database called *ICBHI Challenge 2017*, which is composed of 920 respiratory sounds acquired from 126 participants. The database contains 6898 respiratory cycles that include wheeze, crackles, or a combination of both, as well as cycles without adventitious sounds. In this thesis, the proposed method was evaluated using 575 respiratory cycles containing wheezes, belonging to 45 patients, and 575 normal respiratory cycles, from 44 subjects. The best experimental result reached a precision of 0.880, accuracy of 0.872 and sensitivity of 0.860 for the validation set using the SP method. As well as a precision of 0.666, accuracy of 0.705 and sensitivity of 0.823 for the test set when applying the SST method. When evaluating the validation set, HHS method scored 0.779 for accuracy, 0.790 for sensitivity, 0.767 for specificity, and 0.772 for precision. While the WVD achieved 0.830 accuracy, 0.833 sensitivity, 0.827 specificity and 0.828 precision with the same set. When comparing this performance with those reported in the literature, it was found that the results obtained are at least comparable with those indicated in most state-of-the-art works. We consider that the results of this thesis are promising in detection of respiratory cycles containing wheeze compared to traditional procedures and seem to point to high resolution TFR, especially SST, being a good contender against the current method. SP, and its implementation in a subsequent system can support the diagnosis of obstructive respiratory diseases.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ABREVIATURAS

AI	Inteligencia artificial (*Artificial intelligence*)
CEEMDAN	Descomposición empírica de modos de ensamble completo con ruido adaptativo (*Complete ensemble empirical mode decomposition assisted by noise*)
CNN	Red neuronal convolucional (*Convolutional neural network*)
DNN	Red neuronal profunda (*Deep neural network*)
ECG	Electrocardiografía
EEMD	Descomposición empírica de modos de conjunto (*Ensemble empirical mode decomposition*)
EMD	Descomposición empírica de modos (*Empirical mode decomposition*)
EPOC	Enfermedad pulmonar obstructiva crónica
FFT	Transformada rápida de Fourier (*Fast* Fourier *transform*)
HHS	Espectro de Hilbert-Huang (Hilbert-Huang *spectrum*)
ICBHI	Conferencia internacional en informática biomédica y de la salud (*International Conference on Biomedical and Health Informatics*)
IMF	Modo de oscilación intrínseco (*Intrinsic mode function*)
MI	Información mutua (*Mutual information*)
MM	Mezcla de modos (*Mode mixing*)
SP	Espectrograma (*Spectrogram*)
SR	Sonido Respiratorio
SST	Transformada Synchrosqueezing (Synchrosqueezing *transform*)
STFT	Transformada de Fourier de tiempo corto (*Short-time* Fourier *transform*)

TEWA Análisis de forma de onda expandida en el tiempo (*Time-expanded waveform analysis)*

TF Tiempo-Frecuencia (*Time-Frequency*)

TFR Representación Tiempo-Frecuencia (*Time-Frequency representation*)

WVD Distribución Wigner-Ville (Wigner-Ville *distribution*)

Capítulo 1

Introducción

I.1. Enfermedades respiratorias

Las enfermedades pulmonares se sitúan en el tercer lugar dentro de las diez principales causas de muerte a nivel mundial, de acuerdo con la organización mundial de la salud (OMS). Según cifras de la OMS, el asma afectó a un estimado de 262 millones de personas en el año 2019 y causó 461,000 muertes en el mundo. La mayoría de las muertes relacionadas a esta enfermedad ocurren en países de ingresos bajos y medios-bajos, donde las bajas tasas de diagnóstico y de tratamiento son un desafío. Mientras que la enfermedad pulmonar obstructiva crónica (EPOC) representa la tercera causa de muerte en países de ingresos medios-altos, ocasionando 2.34 millones de defunciones en 2019, como se muestra en la gráfica de la **Fig. 1** [1]. Ambas patologías son enfermedades no transmisibles, obstructivas, crónicas y que comparten síntomas, como el estrechamiento de las vías aéreas, dificultad respiratoria, disminución del flujo espiratorio y aparición de sibilancias en los sonidos respiratorios (SR) [2].

El conjunto de enfermedades broncopulmonares también ha ido en aumento en México [3]. En el año 2019, la OMS reportó la EPOC como la séptima causa de muerte en nuestro país, como indica la **Fig. 2**, con 22.75 fallecimientos por cada 100,000 habitantes [4]. La EPOC está caracterizada por síntomas respiratorios como sibilancias, disnea, tos, expectoración y dentro de sus factores de riesgo se incluyen la exposición prolongada a gases, humo de tabaco, contaminación y uno de los más relevantes que es padecer asma en la infancia [2]. El humo de cigarrillo induce la pérdida acelerada de la función pulmonar, incapacidad respiratoria y muerte anticipada, además de provocar dos síntomas importantes:

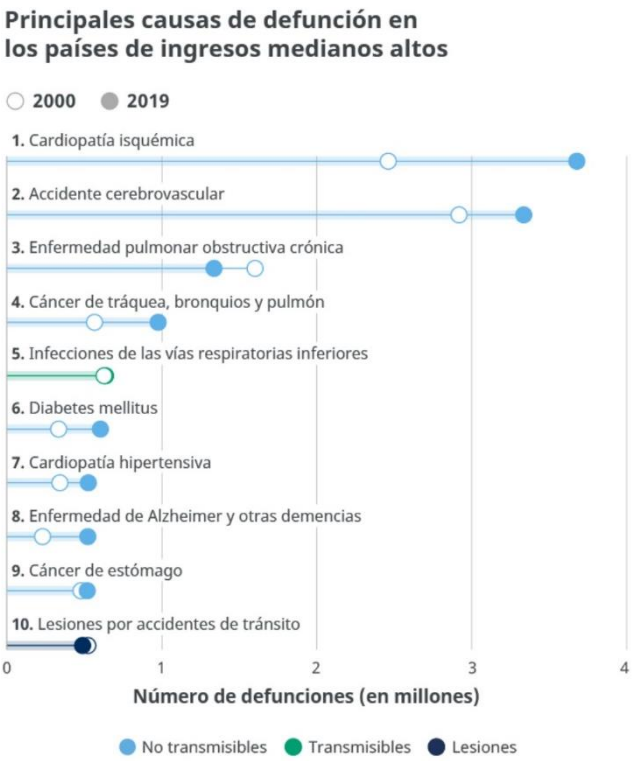

Fig. 1. Principales causas de defunción a nivel mundial [1].

1) aumento en la secreción de moco, deterioro del sistema mucociliar y decremento de las defensas mecánicas presentes en el aparato respiratorio, lo que conduce a aumento de infecciones, tos y expectoración crónica, y 2) lesiones en las vías aéreas pequeñas y en el parénquima pulmonar, lo que reduce el flujo de aire a través de las vías aéreas [3]. Eventualmente, los sacos alveolares aumentan de tamaño forzando al organismo a buscar más oxígeno, induciendo una insuficiencia respiratoria.

El asma es un trastorno respiratorio que aparece en el 3-5% de todas las personas durante algún momento de su vida [5]. Esta enfermedad se caracteriza por inflamación y broncoconstricción, edematosa e inflamatoria, que resulta en episodios de tos y secreciones. Existen episodios agudos y reversibles que pueden desencadenarse debido a infecciones o a inhalación de sustancias contaminantes ambientales o alergenos, como pólenes de plantas, humo de tabaco, smog, entre otros. Mientras que las crisis que ocurren con mayor frecuencia suelen evolucionar a una enfermedad más grave como enfisema pulmonar [3], enfermedad

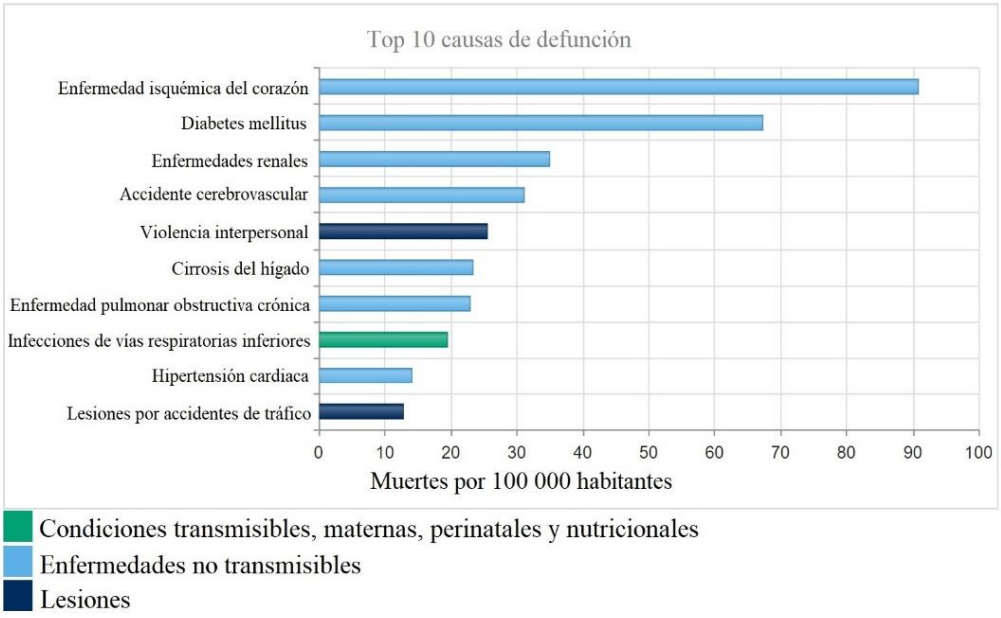

Fig. 2. Diez principales causas de fallecimientos en México en 2019 [4].

en la cual ocurre disminución de la elasticidad en el tejido pulmonar y disminución de la velocidad de flujo espiratorio [6]. Debido al episodio asmático, el diámetro de los bronquiolos disminuye más su tamaño durante la espiración que en la inspiración [5].

Se ha reportado que los hijos de madres que fueron fumadoras durante el embarazo tienen mayor probabilidad de desarrollar asma, particularmente durante los tres primeros años de vida del niño [7]. Además, existen más de 200 sustancias que pueden causar asma laboral [3]. Se sabe que existe una mayor prevalencia de esta enfermedad en mujeres que trabajan realizando limpieza doméstica, que las que no han trabajado en esta área, al estar en contacto con productos de limpieza particularmente en áreas reducidas o poco ventiladas [8].

Existen enfermedades obstructivas crónicas que son generadas por otros mecanismos como: 1) la ocupación del lumen bronquial causada por edema, exceso de secreciones o hipertrofia de glándulas mucosas, e.g., bronquiectasias y fibrosis quística, 2) inflamación y edema de la pared de los bronquios, e.g., bronquitis crónica, o 3) causas externas a las vías aéreas por alteración de la elasticidad pulmonar, e.g., enfisema pulmonar [6].

I.2. Sonidos respiratorios

Los SR son sonidos producidos por la turbulencia de aire a través de las vías aéreas. En sujetos sanos los SR son considerados como ruidos suaves, susurrantes y de baja tonalidad. Cualquier sonido que difiera de estas características es sugestivo de anormalidad o presencia de enfermedades, por lo que se considera que los SR son indicadores relevantes de salud y condiciones patológicas, ya que están directamente relacionados con la fisiología de las vías aéreas, como cambios en el tejido pulmonar y localización de secreciones en las vías respiratorias y los pulmones. Los componentes principales de SR se encuentran en una banda de frecuencias entre 100 Hz y 2,500 Hz [9], [10]. Los SR reflejan cambios dependiendo de la patología presente en las vías aéreas, e.g., la obstrucción bronquial como la que se presenta en el asma aumenta los componentes de alta frecuencia y altas intensidades, mientras que en la broncodilatación la energía se mueve hacia frecuencias más bajas. Además, se ha encontrado que en el asma hay una asociación muy significativa entre el nivel de broncoconstricción y la frecuencia promedio de los SR [9].

Típicamente los SR se clasifican como normales o adventicios. Los sonidos adventicios se encuentran agregados o superpuestos en los SR normales y se pueden clasificar como continuos e.g., sibilancias o *wheezes*, o discontinuos e.g., crepitancias o *crackles*, dependiendo de su duración [9], [11]. Entre los sonidos discontinuos se encuentran las crepitancias, caracterizadas por ser de carácter explosivo y de corta duración, por lo general menor a 20 ms, con un amplio rango de frecuencia (100 – 2,000 Hz) y se suelen presentar durante la inspiración. Se pueden dividir en crepitancias finas o gruesas, las primeras tienen un tono principal de frecuencia de 650 Hz y una duración de 5 ms, mientras que las crepitancias gruesas suelen tener un componente de frecuencia de 350 Hz y duración de 15 ms. Se cree que las crepitancias son originadas por la energía acústica generada por la ecualización de presión o un cambio en estrés elástico, después de una apertura súbita de vías aéreas normalmente cerradas [9]. Este tipo de sonidos adventicios discontinuos está asociada a enfermedades obstructivas como EPOC, bronquitis crónica, neumonía y fibrosis pulmonar.

I.2.1. Sonidos respiratorios normales

Los componentes principales de las señales de SR se encuentran en una banda de frecuencia de entre 100 Hz y 2,500 o hasta 4,000 Hz. El límite inferior se determina para eliminar los componentes de frecuencias bajas como los sonidos musculares, del corazón, movimiento del sensor de adquisición o del paciente y ruido externo de baja frecuencia [12], [13]. El límite superior puede variar dependiendo de la aplicación, si se adquieren señales de vías aéreas superiores se opta por una frecuencia de corte mayor. Además, su contenido en frecuencia debe tomar en cuenta el lugar anatómico donde se realiza la adquisición de la señal (hasta 2 kHz en pecho y hasta 4 kHz en tráquea; para sonidos adventicios en pecho hasta 6 kHz) [9], [14]. La forma de onda resultante de los sonidos normales está desorganizada, es decir, contiene muchas frecuencias diferentes, como se aprecia en la **Fig. 3A**, que corresponde al fononeumograma de un sujeto sano

I.2.2. Sonidos respiratorios adventicios continuos

Dentro de los sonidos adventicios continuos se encuentran las sibilancias, o *wheezes*, los cuales son sonidos continuos, que están sobrepuestos en los SR normales y son más fuertes que éstos, por lo que suelen ser audibles cuando el paciente tiene la boca abierta o mediante auscultación de la laringe, ya que la transmisión de este tipo de sonidos es mejor mediante las vías aéreas que a través de los pulmones hacia la superficie del tórax. Los componentes de altas frecuencias tienden a ser absorbidos por el tejido pulmonar [9].

Las sibilancias están caracterizadas por tener una duración más larga que los sonidos adventicios discontinuos y son definidas por el proyecto CORSA (*computerized respiratory sound analysis*, análisis de SR computarizados) como un sonido agudo, con una duración mayor a 100 ms y un componente dominante de frecuencia por arriba de los 100 Hz [9]. La forma de onda de las sibilancias semeja la de una onda sinusoidal. Un ejemplo se muestra en el panel de la **Fig. 3B**, el cual muestra la señal de SR de un paciente con asma [9].

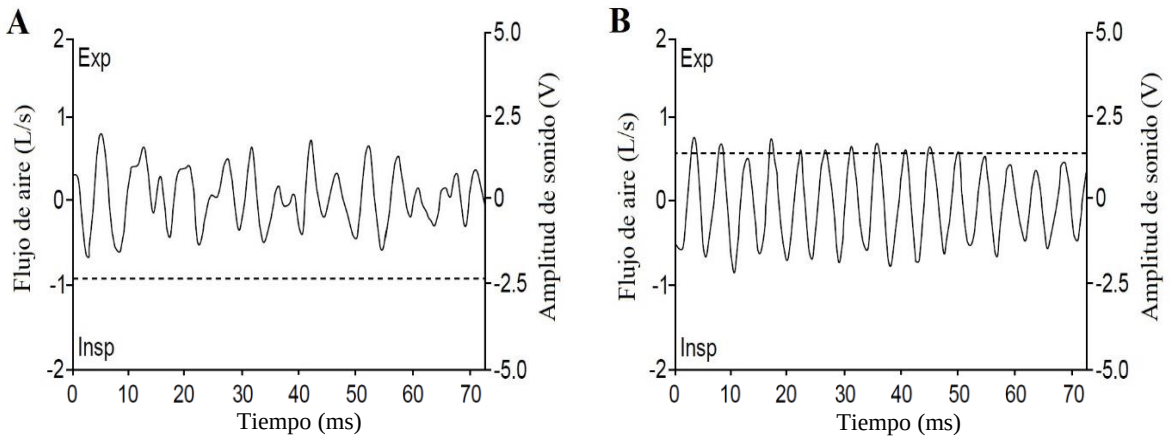

Fig. 3. Fononeumogramas de sonidos respiratorios: A) normal y B) con sibilancias [9].

Las sibilancias son sonidos de carácter musical, su forma de onda contiene uno o más componentes sinusoidales [15], lo que se ve reflejado en el dominio de la frecuencia en el espectro de potencia por medio de picos de distintas frecuencias [10], [16]. Las sibilancias se pueden clasificar de acuerdo con el rango de frecuencias que presentan, siendo: 1) monofónica, donde un solo tono o una sola frecuencia está presente, o 2) polifónica, cuando múltiples frecuencias son percibidas simultáneamente.

I.3. Relevancia clínica de la detección y análisis de las sibilancias

En la práctica clínica, el número de sibilancias, su localización y distribución sobre el pecho, así como su duración, corresponden a la distribución de la patología pulmonar y ofrecen información útil a los médicos para la identificación de las enfermedades respiratorias [9], [12]. La aparición de una sibilancia en un área delimitada indica la presencia de una vía aérea parcialmente obstruida; mientras que múltiples sibilancias, de tono variable, con inicio y fin en tiempos diferentes, sobre ambos pulmones, son indicativos de asma y EPOC [16]. Se ha reportado que en pacientes con asma obstructiva aparece un mayor número de sibilancias que en sujetos sanos al realizar maniobras espiratorias forzadas [17], lo que sugiere que un mayor porcentaje de las vías aéreas son afectadas por la enfermedad. Además, se han estudiado características de los SR encontrados en ambos grupos, sanos y enfermos, encontrando que

las sibilancias polifónicas son más comunes en pacientes con patologías obstructivas, lo que aumenta la probabilidad de que las sibilancias se generen en diversas zonas de las vías respiratorias [18].

Las técnicas de análisis tiempo-frecuencia (*time-frequency*, TF) brindan información certera en la detección y ubicación de características de los SR y ofrecen características importantes en el dominio temporal y espectral, de manera que la combinación de ambas puede describir completamente los eventos de interés, mediante la extracción de sus propiedades principales y el análisis de la variación de su frecuencia a lo largo del tiempo.

I.4. Detección automática de sibilancias

En las últimas décadas, se han realizado esfuerzos para realizar la detección automática de sibilancias y de ciclos respiratorios que las contienen. Estos procedimientos pretenden incrementar la sensibilidad de la detección basada, usualmente, en alguna representación tiempo-frecuencia (*time-frequency representation*, TFR). El enfoque de estos modelos es la extracción de características de la señal y comparación de algoritmos para mejorar el reconocimiento de esas propiedades. El propósito de extraer características con estos métodos es tomar una señal y reducirla a una cantidad pequeña de parámetros o variables para su posterior procesamiento y análisis.

Las primeras aproximaciones utilizaban modelos de clasificación cuyos coeficientes debían ser ajustados empíricamente por el equipo de trabajo, basándose en el conocimiento existente sobre los distintos SR, lo cual resultaba inconveniente, no uniforme o estandarizado, e incrementaba la complejidad y el tiempo de procesamiento de los algoritmos. Estos modelos son capaces de realizar la detección precisa de sibilancias, pero son lentos [15]. Posteriormente, con el desarrollo de la inteligencia artificial (*artificial intelligence*, AI) inició el uso y exploración de técnicas de aprendizaje de máquina o aprendizaje profundo, los cuales brindaron muchas de sus bondades para la detección automática de SR adventicios, e.g., la extracción automática de características de las señales, reducción del tiempo de clasificación, flexibilidad para modificar la estructura del algoritmo, entre otras.

Uno de los primeros métodos modernos de detección fue el implementado por Bahoura y Pelletier en 2003 [19], donde realizaron análisis cepstral para clasificar SR aplicando el método de coeficientes cepstrales de frecuencia de Mel (Mel *frequency cepstral coefficients*, MFCC) en segmentos de señales de audio de SR, los cuales fueron caracterizados por un número reducido de veinte coeficientes. Posteriormente los segmentos fueron clasificados como sonidos normales o sonidos con sibilancias mediante el método de cuantización vectorial. En el sistema de clasificación utilizado, se emplearon dos etapas: una de entrenamiento y una de reconocimiento. En la primera fase, un modelo acústico se construyó para cada tipo de SR y los modelos se guardaron en una base de datos. Mientras que en la fase de reconocimiento, un SR nuevo o desconocido para el algoritmo fue analizado y en la base de datos se buscó el modelo que tuviera mayor coincidencia con el segmento a analizar [19].

Otro trabajo de detección de sibilancias fue el realizado por Lin et al. en el año 2015 [15], quienes propusieron un método de procesamiento de imágenes basadas en análisis TF mediante el uso del espectrograma (SP). Adquirieron registros de SR de sujetos sanos y asmáticos y emplearon el método de orden truncado promedio (*order truncate average*, OTA) para detectar las sibilancias y para determinar sus características musicales mediante una frecuencia fundamental y sus armónicos, obteniendo una sensibilidad de detección de 0.946 y especificidad 1.0 [15].

En un trabajo más reciente, desarrollado por Shethwala et al. en 2021 [20], se emplearon imágenes de espectrograma de Mel. Realizaron una comparación de desempeño de clasificación empleando métodos de aprendizaje de máquina, con diversas redes neuronales pre-entrenadas mediante transferencia de aprendizaje, así como redes neuronales diseñadas por el equipo de trabajo. Finalmente, obtuvieron valores de precisión de 0.64 y exactitud de 0.72 para la clasificación de SR de sujetos sanos y enfermos [20].

Este proyecto de tesis sigue el planteamiento de diversos trabajos actuales que aprovechan las ventajas del aprendizaje de máquina en conjunto con aquellas bondades de las TFR.

I.5. Planteamiento del problema

Los métodos existentes y comúnmente utilizados para realizar la detección automática de sibilancias tienen limitaciones. En el análisis de señales acústicas, como los SR, se suelen emplear técnicas TF que se han estandarizado o cuyo uso se ha normalizado, tal es el caso del SP basado en Fourier. Sin embargo, estas técnicas presentan limitaciones como una baja resolución TF conjunta, que no permite visualizar adecuadamente la naturaleza compleja y variante en el tiempo y en la frecuencia de las señales respiratorias. Además, en algunos de los algoritmos empleados en distintos trabajos no se explora la gran variedad de arquitecturas que ofrecen los métodos de AI, lo que puede resultar en baja especificidad y sensibilidad.

El análisis de sibilancias no es trivial y muchas veces identificarlos a simple vista en una TFR no es sencillo, reflejando diversos factores que influyen sobre su comportamiento como las variaciones anatómicas y fisiológicas que existen entre diferentes individuos e incluso los cambios presentes en el mismo sujeto a lo largo del tiempo [21], i.e., las señales biológicas son complejas y cambiantes en el tiempo. Por lo que implementar un análisis TF clásico y de alta resolución mediante algoritmos de aprendizaje profundo presenta una gran ventaja al permitir describir la energía de la señal en ambas dimensiones TF y realizar un análisis completo de sus propiedades espectrales y variaciones en el tiempo [22], [23]. El detector automático de sibilancias basado en redes neuronales puede alcanzar mayor sensibilidad para la detección correcta de eventos dentro de los SR al identificar y extraer sus características, superando las limitaciones que se presentan al utilizar el método tradicional de auscultación pulmonar que suele ser una técnica subjetiva.

En este trabajo de tesis, se plantea realizar la detección automática de ciclos respiratorios que contienen sibilancias empleando algunas TFR clásicas y TFR de alta resolución. Consideramos que, emplear TFR de alta resolución presentará la ventaja de permitir describir la energía de las sibilancias de manera concentrada en el dominio tiempo-frecuencia conjunto. Además, el análisis del desempeño de los algoritmos contempla el empleo de un conjunto de prueba, lo cual pudiera resultar valioso para evaluar el método propuesto. En la literatura, son escasos los trabajos que utilizan este conjunto, la mayoría de ellos divide sus datos, únicamente, en entrenamiento y validación. De esta manera, usar un conjunto adicional para probar el método propuesto en este trabajo, pudiera brindar una

aproximación de la capacidad de detección de ciclos respiratorios normales y con sibilancias al emplear datos nuevos.

1.6. Objetivo general

Implementar un algoritmo para la detección automática de ciclos respiratorios que contienen sibilancias, empleando representaciones tiempo-frecuencia clásicas y de alta resolución, mediante técnicas de aprendizaje de máquina para su clasificación.

I.7. Objetivos particulares

- Obtener y preprocesar una base de datos pública que contenga SR normales y sibilancias.
- Explorar técnicas TF clásicas y de alta resolución.
- Implementar detectores de ciclos respiratorios con sibilancias basados en las técnicas de TF clásicas y de alta resolución, así como en redes neuronales convolucionales por medio de transferencia de aprendizaje.
- Analizar el desempeño de los algoritmos de detección, incluyendo un conjunto de prueba.

I.8. Estructura general de la tesis

El presente trabajo de tesis se enfoca en implementar un algoritmo de detección de ciclos respiratorios con sibilancias empleando TFR clásicas y TFR de alta resolución, que resulte adecuado para analizar señales acústicas producidas por la respiración en sujetos sanos y enfermos. La organización del resto de capítulos de esta tesis se describe a continuación.

El **capítulo 2** aborda la historia y antecedentes de la auscultación pulmonar, sus inconvenientes iniciales, la evolución de la técnica de adquisición, exploración y

clasificación de algunos SR, en particular de las sibilancias. Además, se plantean métodos de detección de sibilancias basados en análisis temporal, espectral y TF, mencionando esfuerzos realizados en cada una de estas áreas.

En el **capítulo 3** se plantea la importancia del análisis de la señal en los dominios del tiempo y la frecuencia para obtener sus diversas características, así como la relevancia del análisis TF conjunto que ofrece una descripción completa del comportamiento de la señal. Además, se plantean las cuatro técnicas TF a evaluar: la distribución de Wigner-Ville, el espectrograma, el espectro de Hilbert-Huang y la transformada Synchrosqueezing. Para cada técnica, se discuten sus principales características, ventajas y desventajas. Finalmente, se aborda la definición e historia de la inteligencia artificial, las redes neuronales y la técnica de transferencia de aprendizaje.

Posteriormente, en el **capítulo 4** se aborda la metodología empleada, estableciendo la simulación de sibilancias sintéticas, así como la creación de su TFR ideal teórica a emplear como referencia en la evaluación de las técnicas mencionadas en el capítulo 3, así como el método de selección de sus parámetros. También, se describe la base de datos de acceso público empleada en este trabajo de tesis. Se menciona el preprocesamiento de datos, las características de la red neuronal empleada y los índices de desempeño para evaluarla.

En el **capítulo 5** se presentan los resultados obtenidos y su discusión, al procesar las señales simuladas y las de SR obtenidas de la base de datos con las técnicas TF, así como los valores de los índices de desempeño alcanzados por el detector para cada TFR, y los resultados obtenidos en los conjuntos de prueba.

Finalmente, el **capítulo 6** corresponde a las conclusiones obtenidas de este trabajo de tesis, sus limitaciones y trabajo a futuro.

Adicionalmente, en el **Anexo** se presenta el artículo titulado “Evaluación de un detector automático de ciclos respiratorios con sibilancias basado en CNN y representaciones tiempo-frecuencia clásicas y de alta resolución” (*Assessment of an automated wheezing respiratory cycle identification based on CNN and classical and high-resolution time-frequency representations)*. Este trabajo fue presentado en la 1er Conferencia Internacional

de la IEEE de Ingeniería en Medicina y Biología en Latinoamérica 2023 (IEEE *1st Engineering in Medicine and Biology Conference in Latin America* 2023).

Capítulo 2

Antecedentes

El estudio de las enfermedades pulmonares tiene su origen hace varios siglos, a lo largo de los cuales han existido diversas personalidades y métodos para conocer e identificarlas. René-Théophile-Hyacinthe Laënnec (1781-1826) fue uno de los más grandes clínicos de todos los tiempos. Realizó contribuciones importantes en la explicación de padecimientos y lesiones patológicas de enfermedades del tórax, e.g., el enfisema, la bronquiectasia y la tuberculosis, y además inventó y propulsó el uso del estetoscopio. Anteriormente, la técnica para estudiar sonidos pulmonares y cardiacos consistía en aproximar el oído a la caja torácica del paciente, sin embargo, esta técnica generaba muchos inconvenientes, como la percepción de sonidos o información imprecisos, así como la incomodidad o el difícil acceso a los sonidos de interés en personas con mayor masa corporal. Laënnec tuvo la idea de emplear un rollo de papel y, posteriormente un cilindro de madera, **Fig. 4**, instrumentos que le permitían distinguir sonidos que antes no había escuchado [3]. Fue mediante la recopilación de la nueva información que Laënnec comenzó a determinar el cuadro clínico de varias enfermedades [24]. Posteriormente, su estetoscopio fue perfeccionado al pasar de un diseño que permitía la auscultación con un solo oído a otro que permite emplear ambos oídos, tal como se utiliza en la actualidad.

La técnica de auscultación pulmonar es una de las herramientas comúnmente empleadas por los médicos para el diagnóstico de padecimientos respiratorios, e.g., asma y EPOC, y es realizada con un estetoscopio mecánico convencional. Este método es rápido, no invasivo y permite recopilar evidencias que posteriormente se confirman mediante técnicas exploratorias, como la inspección de tórax [6], u otras técnicas más especializadas e.g., radiografía de tórax, pletismografía o espirometría para valorar el daño o funcionamiento de

Fig. 4. Cilindro de madera creado por Laënnec.

los pulmones [3]. Sin embargo, esta técnica tiene limitaciones, e.g., es un proceso subjetivo que depende de la capacidad de audición de quien lo realice, de su experiencia y habilidad para diferenciar sonidos y patrones, y no produce mediciones cuantitativas, ni monitorización por largo tiempo ni correlación entre los SR y otras señales fisiológicas [11].

Ha sido mediante avances producidos en los últimos 50 años que, gracias a métodos computarizados, se han vencido algunas de las limitaciones de la técnica de auscultación tradicional. Algunos de estos logros son: 1) la cuantificación de cambios de sonidos pulmonares, 2) la correlación con otras señales fisiológicas al medirse simultáneamente, e.g., junto con la señal cardiaca o la señal de flujo respiratorio, 3) la evolución de los sistemas de registro electrónicos mediante el uso de micrófonos que simulan estetoscopios convencionales, con mayor resolución de información y amplia variedad de técnicas de procesamiento de señales, y 4) la realización de registros permanentes y de representaciones gráficas que ayudan al diagnóstico, que permiten tener una mejor visualización de los SR y los episodios pulmonares [11]. Así, en las últimas décadas se han desarrollado métodos computarizados de detección de sonidos respiratorios adventicios y otros eventos presentes en registros de SR.

II.1. Métodos de detección de sibilancias

Los sistemas automáticos computarizados de SR ofrecen ventajas sobre la técnica de auscultación tradicional, ya que ofrecen una detección objetiva y un diagnóstico de sonidos patológicos adventicios, en contraste con la experiencia profesional médica que es subjetiva. La mayoría de los sistemas de detección automática para clasificación de SR adventicios comprenden dos partes: 1) extracción de características relevantes de la señal, y 2) las características extraídas son usadas para detectar o clasificar eventos de SR adventicios, i.e. crepitancias y sibilancias [21]. Dependiendo del enfoque de los sistemas los podemos clasificar en tres tipos principales: 1) análisis en el dominio del tiempo, 2) análisis en el dominio de la frecuencia y 3) análisis en el dominio TF conjunto [15]. A continuación, se mencionan sus ventajas y desventajas frente a otros métodos, así como algunos estudios representativos de cada tipo.

II.1.1. Análisis en el dominio del tiempo

El análisis temporal comenzó a implementarse a finales de la década de 1960 para observar el intervalo de tiempo entre eventos consecutivos y su intensidad relativa. En el estudio de los SR, una forma de visualización de estos consiste en desplegar simultáneamente la señal de amplitud de sonido y la señal de flujo respiratorio como función del tiempo, también llamado fononeumograma. Al contar con ambas representaciones resulta sencillo relacionar e identificar la forma de onda de los eventos respiratorios.

En la década de 1970, Murphy et al. caracterizaron sonidos respiratorios al analizar las gráficas amplitud-tiempo de registros de fenómenos auscultatorios [25]. Propusieron el método de análisis de forma de onda expandida en el tiempo (*time-expanded waveform analysis*, TEWA), el cual provee visualizaciones reproducibles que permiten documentar las características de los diversos sonidos respiratorios y pulmonares, como las crepitancias, ronquidos, sibilancias, sonidos normales, e incrementa la utilidad diagnóstica del análisis de sonidos respiratorios [25].

En la **Fig. 5** se muestran gráficas de diferentes sonidos respiratorios, donde en la columna izquierda aparecen los sonidos en una escala de tiempo convencional, para abarcar un ciclo respiratorio, mientras que en la columna derecha se observan los mismos sonidos en su escala de tiempo expandido, los cuales revelan patrones distintivos que no son observables a escala convencional [25].

Al aplicar el método TEWA utilizando señales digitales computarizadas, Murphy et al. obtuvieron una resolución temporal adecuada, lo que permitió observar las formas de onda expandidas en el tiempo de diversos SR típicos y lograron detectar patrones significativos que permitieron distinguir visualmente cada tipo de sonido. Esta diferenciación no habría sido considerada confiable de haberse realizado a partir de las formas de onda registradas por métodos comunes a velocidades convencionales [25]. En el análisis TEWA, los SR normales son observados con una forma irregular y no cuentan con patrones repetitivos o característicos, mientras que los sonidos adventicios continuos presentan una forma de onda periódica sinusoidal [26]. Los ronquidos aparecen como patrones de sonido repetitivos con

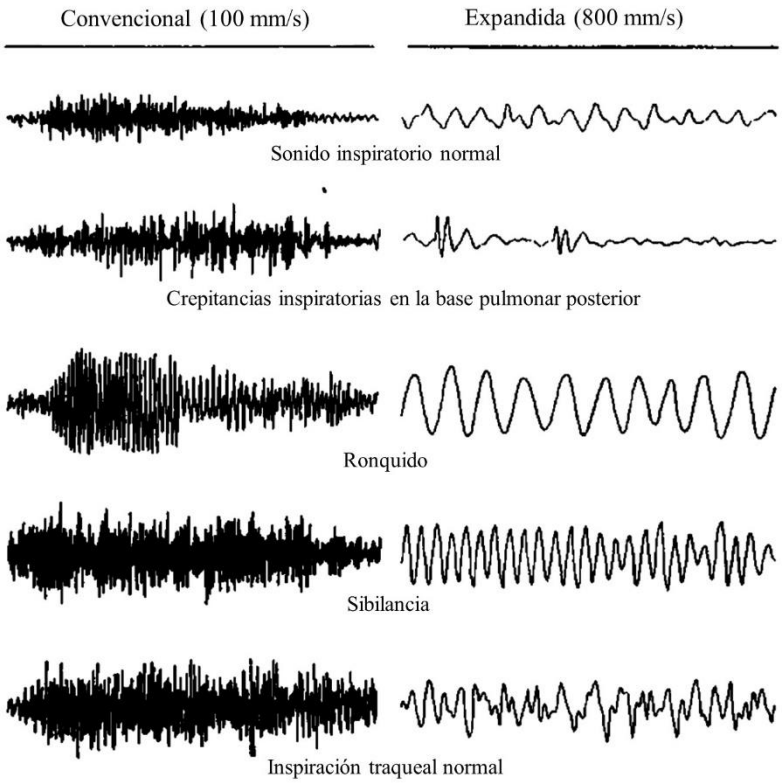

Fig. 5. Análisis TEWA de distintos SR [23].

duración usualmente mayor a 250 ms, y las crepitancias presentan una corta deflexión repentina seguida por desviaciones de amplitud mayor. Esto también permitió clasificar los SR adventicios como continuos y discontinuos.

II.1.2. Análisis en el dominio de la frecuencia

En el análisis espectral, las sibilancias son vistas como picos estrechos en el espectro de potencia, generalmente debajo de 2,000 Hz. Estos algoritmos son rápidos y sencillos, pero no son confiables ni sensibles [15].

Shabtai-Mussih et al. [27], y J. A. Fiz et al. [17] aplicaron un algoritmo de detección de picos en el dominio de la frecuencia para detectar sibilancias; comparando el contenido espectral de las señales, mediante el cálculo de la transformada rápida de Fourier (*fast Fourier transform*, FFT) de segmentos de sonidos, y evaluando la presencia de picos de potencia previamente asociados con sibilancias. Shabtai-Mussih et al. [27] realizaron una selección manual de los segmentos más prominentes, angostos y claramente separados de otros picos. Posteriormente, emplearon una calificación para indicar la probabilidad de que un pico en el espectro de potencia representara una sibilancia, de acuerdo con criterios establecidos por el equipo, registrando aquellos que obtuvieron las calificaciones más altas. Sin embargo, este método presenta dificultades, como la selección del tamaño de segmentos a analizar de la señal respiratoria, la falta de un criterio homogéneo para determinar la potencia de las sibilancias y de los sonidos pulmonares normales, la variabilidad de la frecuencia dominante de las sibilancias, entre otros.

En el estudio de J. A. Fiz et al. [17] realizaron la evaluación de los picos, obtenidos mediante FFT, de acuerdo con ocho criterios establecidos empíricamente por el equipo, respetando el criterio establecido por el grupo de trabajo europeo CORSA referente al componente de frecuencia mayor a 100 Hz. Posteriormente, realizaron un agrupamiento de picos siguiendo cuatro puntos acordados por el equipo de trabajo, buscando aquellas agrupaciones que cumplieran otra característica importante sobre la duración de las sibilancias. Estos métodos representan la dificultad para establecer criterios homogéneos

para todas las sibilancias, ya que la elección de los criterios a evaluar es distinta para cada estudio, siendo determinados por cada grupo de trabajo dependiendo de las señales obtenidas.

II.1.3. Análisis tiempo-frecuencia

Para lograr mayor sensibilidad y desempeño de detección se incluye en estos métodos un conjunto de criterios temporales y espectrales de las sibilancias en el dominio TF. Estos criterios pertenecen a la duración, el rango de tono y la magnitud de las sibilancias en la representación TF, e.g., obtenida a través del análisis del SP [15]. El objetivo de este tipo de métodos es descomponer la señal en funciones base y describir la energía de la señal en la dimensión del tiempo y la frecuencia para realizar un análisis completo de sus propiedades.

Las primeras aproximaciones TF para identificar sibilancias en los SR, como el realizado por Pasterkamp et al [13]. en la década de 1980, compararon el sonograma (espectrograma) de SR normales y adventicios, obtenido por medio del ventaneo y de la FFT. Adquirieron simultáneamente dos señales fisiológicas adicionales como referencias de eventos temporales, la señal de electrocardiografía (ECG) y la señal de flujo o volumen respiratorio. En la **Fig. 6** se muestra el fononeumograma de un paciente con asma y sibilancias posterior a realizar ejercicio, se observan sibilancias polifónicas durante la inspiración y la espiración. En la parte superior se muestra la señal de flujo respiratorio en color amarillo, y la señal de SR original debajo de ella, en color naranja. Mientras que, en la parte superior derecha se muestra la forma de onda del segmento de 100 ms de duración marcado en 6.7 s, perteneciente a una espiración, así como en la parte inferior derecha el espectro de potencia del mismo segmento [13]. En el respirosonograma, los sonidos traqueales alcanzaron los 1800 Hz, y las sibilancias aparecen como concentraciones de alta intensidad de sonido a distintas frecuencias, representadas como líneas cuasi-horizontales en color rojo, tanto en la inspiración como en la espiración. Es importante resaltar que estas líneas cumplen las características de duración u contenido en frecuencia establecidas por el proyecto CORSA [26]. Cabe mencionar que no se realizó una clasificación automática de

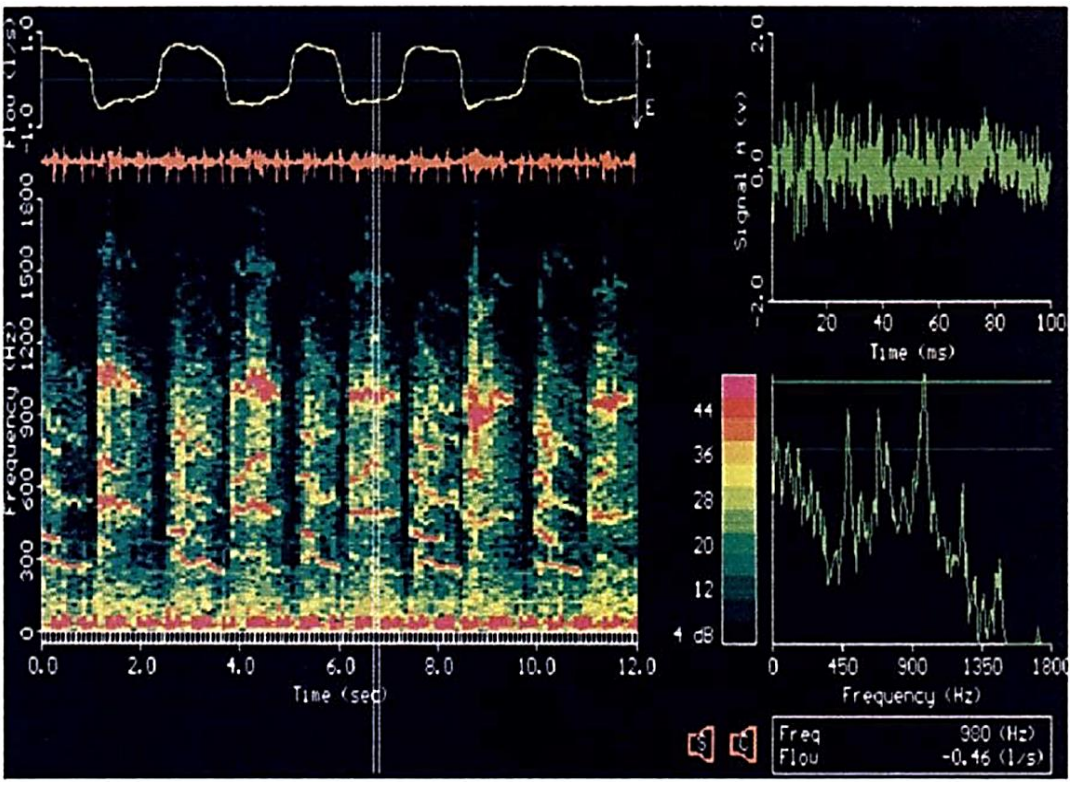

Fig. 6. Imagen de fononeumograma de paciente con asma [13].

sibilancias, sin embargo, este trabajo resultó pionero al complementar la información visual de la forma de onda y del espectro de los SR con su representación TF conjunta [13].

Taplidou et al. [28] emplearon un método de detección de sibilancias basado en la transformada *wavelet* continua (*continuous wavelet transform-based wheezing-episode detector*, CWT-WED), con el cual obtuvieron la ubicación de las sibilancias dentro del ciclo respiratorio y lograron separar la señal de sibilancia del sonido pulmonar adquirido. Obtuvieron el escalograma (magnitud al cuadrado de la transformada *wavelet* continua) destacando la ubicación y características energéticas de la señal, mediante la representación de picos aislados [28] [29].

Se han reportado esfuerzos que emplean alguna TFR para la detección y clasificación automática de sibilancias empleando la base de datos del reto 2017 de la *International Conference on Biomedical and Health Informatics* (ICBHI *Challenge*). Por ejemplo, uno de los primeros trabajos que emplearon esta base de datos y participaron en el reto de clasificación, fue realizado por Jakovljević et al. [30], quienes presentaron un método basado

en los modelos ocultos de Markov (*hidden Markov model*) en combinación con modelos de mezcla Gaussianos (*Gaussian mixture model*) para realizar clasificación de SR en sonidos normales, sibilancias y crepitancias. Como características de entrada al modelo emplearon los MFCC a partir de ventanas de 30 ms de las señales, las cuales fueron preprocesadas aplicando filtros, supresión de ruido mediante sustracción espectral y remuestreo de las señales. Finalmente obtuvieron una exactitud de clasificación general de 0.395 [30].

En otro estudio, Kochetov et al. [31], propusieron una arquitectura de red neuronal recurrente con enmascaramiento de ruido (*noise masking recurrent neural network*, NMRNN) para la clasificación de sonidos pulmonares. Su modelo aprendió a extraer cuadros importantes de tipo respiratorio sin ruido redundante, i.e., el modelo fue entrenado sin realizar una etapa de preprocesamiento para separar los ciclos respiratorios. A partir de esa información realizaron el entrenamiento para clasificar los SR en las cuatro categorías que contiene la base de datos ICBHI: 1) SR normales, 2) sonidos con sibilancias, 3) sonidos con crepitancias, y 4) SR con ambos sonidos adventicios, sibilancias y crepitancias. Mediante el uso de este modelo obtuvieron una sensibilidad máxima de 0.56 y especificidad de 0.73 en la clasificación [31].

Otros esfuerzos reportados en la literatura han empleado imágenes de espectrograma de Mel como entrada a redes neuronales para su clasificación, como el trabajo realizado por Demir et al. [32], quienes implementaron un modelo que incluye una red neuronal convolucional (*convolutional neural network*, CNN) de diseño propio, pre-entrenada con un conjunto particular de sus imágenes de SP, para la extracción de características profundas. Dentro de la arquitectura de la red se agregó una estructura paralela que emplea técnicas de agrupación máxima, *max-pooling*, y agrupación basada en promedio, *average-pooling*, buscando reducir las dimensiones espaciales de los mapas de características y optimizar el desempeño de clasificación. Posteriormente, las características extraídas fueron utilizadas como entrada a un clasificador de análisis discriminante lineal (*linear discriminant analysis*, LDA) utilizando el método de conjuntos de subespacios aleatorios (*random subspace ensembles*, RSE). Empleando la misma base de datos ICBHI *Challenge* 2017, obtuvieron una sensibilidad de 0.40, especificidad de 0.99, precisión de 0.86 y una calificación F-score de

0.55 para la clase de sibilancias, mientras que la exactitud promedio para las cuatro clases fue de 71.15% [32].

El trabajo realizado por Shethwala et al. [20], también emplea imágenes de espectrograma de Mel, realizando una comparación entre diversas redes pre-entrenadas, como VGG-16 y ResNet-50, mediante transferencia de aprendizaje, así como una CNN con una estructura diseñada por el equipo, obteniendo valores de precisión de 0.64 y exactitud de 0.72 para la clasificación en las cuatro clases originales de la base de datos ICBHI [20]. Otro trabajo que utiliza este tipo de imágenes es el implementado por Gairola et al. [33], quienes emplean un modelo de red neuronal profunda (*deep neural network*, DNN) y diversas técnicas de aumento de datos como: 1) el aumento basado en concatenación, que genera muestras nuevas de una clase juntando dos segmentos de la misma clase, 2) el relleno inteligente para unificar la longitud de las señales de audio, alargando los ciclos respiratorios cortos, 3) corte de regiones en blanco, para eliminar secciones de la imagen de SP donde no hay información, y 4) ajuste específico del dispositivo, al entrenar una red para cada instrumento de adquisición con el que se obtuvieron los registros de audio. Mediante este método obtuvieron una sensibilidad de 53.7% y especificidad de 83.3% al realizar el entrenamiento para las cuatro clases originales de la base de datos ICBHI [33].

Existen trabajos como el realizado por Chen et al. [23], quienes emplearon la base de datos ICBHI realizando una clasificación de SR utilizando la transformada-S optimizada (*optimized S-transform*, OST) y redes residuales profundas (*deep residual network*, ResNets), para obtener una clasificación en tres categorías: 1) sibilancias, 2) crepitancias, y 3) SR normales. Las señales de audio fueron separadas por ciclos respiratorios, sin realizar preprocesamiento, para aplicar la OST a cada segmento y obtener la imagen TF, que incluye sonidos cardiacos y otros artefactos. Los resultados mostraron una exactitud de 98.79%, sensibilidad de 96.27% y especificidad de 100% [23].

Capítulo 3

Marco teórico

III.1. Señal en los dominios del tiempo y la frecuencia

El análisis de señales radica en el estudio y la caracterización de sus propiedades elementales. A lo largo del tiempo, se ha desarrollado a la par del descubrimiento de señales fundamentales presentes en la naturaleza, como el campo eléctrico, ondas de sonido y corrientes eléctricas. Generalmente, una señal es una función que involucra muchas variables, como el tiempo, el espacio o la frecuencia, entre otras [34]. La descripción de una señal en el campo conjunto tiempo-frecuencia (TF) obtiene su fundamento y motivación a partir del análisis clásico de la señal en el tiempo y la señal en la frecuencia.

El tiempo es fundamental en una señal ya que, al avanzar continua e irreversiblemente, permite visualizar la secuencia de eventos y variaciones de la misma, y describe valores de la señal siguiendo un orden temporal [35]. Sin embargo, para tener un mayor entendimiento del comportamiento de la señal resulta útil estudiarla en una representación distinta. Esto se logra al expandir la señal en un conjunto completo de funciones que, desde un punto de vista matemático, puede ser realizada de infinitas maneras. Lo que caracteriza a una buena representación es la facilidad con que se pueden comprender ciertas características de la señal, ya que dicha representación muestra una cantidad física importante en la naturaleza de la señal. Además del tiempo, la representación más importante es la frecuencia [34].

Jean-Baptiste-Joseph Fourier (1768-1830) fue un matemático y físico francés que desarrolló las matemáticas dedicadas a la representación en la frecuencia de una señal, y en

el año 1807 creó las matemáticas para manejar discontinuidades. El análisis de Fourier que indica que una función arbitraria, continua o discontinua, puede ser expresada como la suma de funciones continuas [36] revolucionó la matemática, la ciencia y la ingeniería, al ser una de las más grandes innovaciones en estas áreas del conocimiento [34], [37].

Existen importantes razones para el análisis espectral o de frecuencia. Primero, al analizar una señal o forma de onda aprendemos algo sobre la fuente de la señal, procedimiento que ha brindado conocimiento sobre la composición de casi todo lo que conocemos. Segundo, la propagación de ondas a través de un medio, por lo general, depende de la frecuencia, ya que ondas de distintas frecuencias se propagan a distintas velocidades. Tercero, la descomposición espectral facilita nuestro entendimiento de las señales. En general, las señales tienden a ser desordenadas, sin embargo, el desorden es en realidad la superposición de señales sinusoidales, lo que resulta más sencillo de comprender y caracterizar al simplificarlo en funciones menos complejas. Finalmente, el análisis de Fourier es una poderosa herramienta matemática para solucionar ecuaciones diferenciales ordinarias y parciales [34].

El análisis espectral, como la densidad espectral de potencia obtenida mediante la transformada de Fourier, indica las frecuencias existentes a lo largo de la duración total de una señal. Sin embargo, tiene una limitación, ya que no brinda información sobre el momento exacto en que cada frecuencia se presenta [34]. Si oscilaciones en diferentes tiempos tienen la misma frecuencia, o aquellas que ocurren en el mismo instante se traslapan con otras de diferentes longitudes de onda, esa información crucial, importante para entender la señal, se pierde en la transformación de Fourier [35], por lo que un acercamiento sobre cómo las frecuencias cambian a través del tiempo es necesario. Esta razón ha dado pie a desarrollar técnicas que se revisarán más adelante.

III.2. Importancia del análisis tiempo-frecuencia

El análisis individual de una señal en el dominio del tiempo o en el dominio de la frecuencia por sí solas no ofrecen una descripción completa del comportamiento de la señal, ya que su densidad de energía o potencia en el tiempo y su correspondiente densidad en la frecuencia,

al estar separadas, no son suficientes para describir la relación que existe entre ambas dimensiones. En la vida diaria empleamos señales cuyas frecuencias cambian a través del tiempo, tal como el tono que utilizamos al hablar para dar énfasis y distintos significados a las palabras o los tonos musicales que comprende una canción, ambos son ejemplos que ilustran la necesidad de una TFR [34].

El objetivo del análisis TF consiste en emplear una función que describa la potencia de una señal simultáneamente en los dominios temporal y espectral para realizar un análisis completo de sus propiedades. Este análisis incluye criterios conjuntos en el dominio TF, e.g., la duración, el rango de tono y la magnitud de una señal o de sus componentes. Generalmente, el plano TF descompone la señal en segmentos definidos por el intervalo de tiempo en el que ocurren y las frecuencias en las que fluctúan sus componentes oscilatorios [35].

Una de las principales ventajas de este análisis es que permite determinar si una señal contiene un solo componente o más de uno. Una señal multicomponente tiene regiones bien delimitadas dentro del plano TF. Existen diversas razones para la variación de la frecuencia en el tiempo, pero se pueden resumir en dos cuestiones principales [34]: 1) la producción de una frecuencia particular depende de parámetros físicos que pueden cambiar a lo largo del tiempo, e.g., la longitud y tensión de una cuerda de violín que se hace resonar, y 2) la propagación de ondas en un medio usualmente depende de la frecuencia, e.g., la luz y el sonido se propagan en el aire sin cambios en las frecuencias que nuestros ojos y oídos son capaces de detectar, de manera que dos observadores distintos pueden visualizar los mismos colores y escuchar los mismos sonidos provenientes de una sola fuente. De forma similar, el contenido en frecuencia de las sibilancias varía dependiendo de diversos factores, entre los que se encuentran parámetros físicos, e.g., la cantidad, localización y severidad de las obstrucciones bronquiales presentes en las vías aéreas del paciente, que a su vez dependen de la patología que se padezca [9] [18]. Además, la frecuencia de las sibilancias puede ser percibida de maneras distintas, dependiendo de la atenuación producida por los diferentes tejidos o medios que constituyen el cuerpo, por los que tiene que atravesar la onda de sonido. Existen diversos factores que afectan la atenuación de los sonidos, en particular de las sibilancias, e.g., la viscosidad del tejido, la distancia que viaja la onda a través del medio, el tipo de tejido, la edad del sujeto, entre otros [38].

III.3. Representaciones tiempo-frecuencia

Para lograr una mayor sensibilidad y desempeño en la detección de las sibilancias se puede incluir, en los modelos de clasificación, un conjunto de criterios conjuntos en el dominio TF, como se mencionó en la sección anterior. Estos criterios inherentes a las sibilancias se pueden observar fácilmente mediante la representación TF, e.g., obtenida a través del SP [15]. El objetivo de este tipo de métodos consiste en visualizar simultáneamente una señal en los dominios temporal y espectral para realizar un análisis completo de sus propiedades.

Este análisis se puede realizar en el plano TF, el cual consiste en un arreglo bidimensional de frecuencias de la señal *vs* el tiempo, cada uno indicado en un eje distinto. Por ejemplo, para realizar la descomposición de una señal en una representación TF, el dominio temporal se segmenta en intervalos del mismo tamaño, para posteriormente realizar un análisis de Fourier en cada segmento [35]. El resultado es una señal transformada bidimensional que cuenta con una variable independiente de frecuencia y una variable independiente de tiempo.

Los primeros trabajos enfocados en el análisis TF empleaban el SP, definido como la magnitud al cuadrado de la transformada de Fourier de tiempo corto (*short time* Fourier *transform*, STFT), para visualizar las sibilancias. Sin embargo, es bien sabido que el SP presenta un compromiso inherente entre su resolución temporal y su resolución espectral debido al empleo del ventaneo para producir la TFR. Como alternativa al SP, se han desarrollado técnicas TF de alta resolución, que descomponen la señal en funciones base. Una de ellas es la transformada *Synchrosqueezing* (*Synchrosqueezing transform*, SST), que considera la energía regional en el plano local TF durante la descomposición y proporciona una representación avanzada de las características TF con componentes inseparables [14]. Otra técnica TF de alta resolución es la correspondiente al espectro de Hilbert-Huang (*Hilbert-Huang spectrum*, HHS), desarrollado para analizar datos no lineales y no estacionarios, permitiendo localizar eventos en el tiempo, así como en la frecuencia a partir de la descomposición de la señal bajo análisis en sus modos de oscilación intrínsecos (*intrinsic mode functions*, IMF) mediante el algoritmo de descomposición empírica de modos (*empirical mode decomposition*, EMD) [39]. En este trabajo, se propone emplear como cuarto método TF a la distribución de Wigner-Ville (Wigner-Ville *distribution*, WVD), la

cual emplea la transformada de Fourier del producto de una señal con su complejo conjugado [34]. Existen otros métodos TF empleados en análisis de señales, sin embargo, se eligieron estas TFR por sus características, al emplear el SP como estándar de comparación, al ser la técnica clásica empleada para análisis TF. Además, se planteó la exploración de técnicas TF de alta resolución: 1) SST ya que, hasta nuestro conocimiento, ha sido empleada en análisis de otro tipo de señales, pero no para realizar detección automática de sibilancias y resulta interesante su similitud con el SP, ya que ambas emplean el análisis de Fourier, y 2) HHS que ha dado buenos resultados con otro tipo de señales fisiológicas donde extraen la señal respiratoria. Finalmente, la exploración de la WVD permitió conocer la capacidad de las redes neuronales para clasificar imágenes que para la visualización humana presenta términos extras que pueden causar confusión al clasificarlos manualmente. No se agregaron más métodos por el tiempo limitado del programa de maestría, sin embargo, puede resultar valioso explorar su implementación.

En la siguiente sección se describen algunas características de las TFR empleadas en este trabajo de tesis.

III.3.1. Distribución Wigner-Ville

La WVD ha sido investigada por diversos autores como una herramienta para realizar análisis TF de distintas señales. Esta transformada TFR es considerada como la más antigua e importante representación cuadrática, fue desarrollada en 1932 por el posterior ganador del Premio Nobel de física, Eugene P. Wigner, quien la aplicó en mecánica cuántica donde desarrolló parte de la física del estado sólido. Dieciséis años después, el investigador en comunicación J. Ville, introdujo esta transformada al procesamiento de señales con el propósito de esclarecer el concepto de frecuencia instantánea [35].

La WVD presenta la transformada de Fourier del producto de una señal con su complejo conjugado, asemejándose al cálculo de la densidad espectral de potencia. Si tenemos una señal analógica $x(t)$, su WVD (**Ec. 1**) escrita como $X_{WVD}(t,\omega)$, es la transformada de Fourier del producto $x(t+\tau/2)x^{*}(t-\tau/2)$:

$$X_{WVD}(t,\omega) = \int_{-\infty}^{\infty} x\left(t+\frac{\tau}{2}\right)\, x^*\left(t-\frac{\tau}{2}\right) e^{-j\omega\tau} d\tau \quad (1)$$

donde el operador * indica el complejo conjugado de la señal. Para obtener la WVD en un tiempo particular se incluyen elementos del producto de la señal en un tiempo pasado $x\left(t+\frac{\tau}{2}\right)$ multiplicado por la señal en un tiempo futuro $x^*\left(t-\frac{\tau}{2}\right)$, donde el tiempo en el futuro es igual al tiempo en el pasado. Mientras que, en términos de su espectro, $X_{wVD}(t,\omega)$ se puede definir como:

$$X_{WVD}(t,\omega) = \int_{-\infty}^{\infty} X^*\left(\omega+\frac{\theta}{2}\right)\, X\left(\omega-\frac{\theta}{2}\right) e^{-jt\theta} d\theta \quad (2)$$

donde $X^*\left(\omega+\frac{\theta}{2}\right)$ denota la señal en una frecuencia anterior, la cual se multiplica por la señal en una frecuencia posterior $X\left(\omega-\frac{\theta}{2}\right)$. La WVD es parte de la clase bilineal de TFR ya que la señal, o su espectro, aparece dos veces en el cálculo. Por lo que, para determinar las propiedades de la WVD en un tiempo t podemos imaginar que doblamos la señal en ese tiempo, y sobreponemos la parte izquierda de la señal sobre la parte derecha para buscar si existe algún traslape entre ambas mitades. Si existe el traslape, esas propiedades están presentes en el tiempo t [34]. Ya que la característica de esta distribución es idéntica en el dominio del tiempo como de la frecuencia, esta propiedad también ocurrirá en el dominio espectral. Otra característica importante de la WVD es que es no local, ya que atribuye el mismo peso tanto a tiempos cercanos como a tiempos lejanos.

Esta distribución emplea la señal analítica, la cual se puede definir en el dominio del tiempo mediante la transformada de Hilbert de la señal real, mientras que en el dominio de la frecuencia se caracteriza por poseer un espectro definido únicamente para frecuencias positivas, siendo este el doble del espectro de la señal real, sobrellevando así la dificultad de interpretar las frecuencias negativas inherentes a las señales reales [35].

Algunas de sus propiedades son: 1) que es no positiva, 2) siempre es real, incluso si la señal es compleja, y 3) para espectros simétricos la WVD es simétrica en el dominio de la

frecuencia, así como para formas de onda simétricas la WVD es simétrica en el dominio temporal. Una de las propiedades importantes de la WVD, así como del SP y cualquier otro miembro de la clase general de TFR bilineales e invariantes a corrimientos en el tiempo y la frecuencia, es que la suma de dos señales no es la suma de las WVD de cada señal. La WVD cruzada es compleja y contiene un par de términos llamados auto-componentes y otro término restante llamado término cruzado [34]. El término cruzado ocurre en la distancia media entre dos auto-componentes en el espacio TF y pueden poseer una amplitud mayor al producto de las magnitudes de los auto-términos. Por lo tanto, la presencia de los términos cruzados puede dificultar enormemente la interpretación de la WVD [40]. Estos términos pueden minimizarse al realizar un suavizado, sin embargo, este procedimiento destruye otras propiedades deseables de la WVD [34].

Quizás por la gran cantidad de términos cruzados en la WVD, su empleo para tareas de clasificación ha sido generalmente ignorada en el campo de SR. Sin embargo, en este trabajo de tesis, consideramos que, al emplear una red neuronal convolucional para realizar la clasificación de SR, la red pudiera ser capaz de extraer las características importantes de la señal, incluso al incluir los términos cruzados, como información importante de cada ciclo respiratorio para lograr buenos resultados de clasificación. Cabe mencionar que, a pesar de sus dificultades, la WVD es única para cada señal y no requiere de la selección de parámetros tales como la longitud o tipo de ventana o la cantidad de componentes resultantes de una descomposición.

III.3.2. Espectrograma

El análisis de Fourier induce la idea de que las frecuencias no están localizadas en el tiempo. A pesar de que esta técnica resalta los diferentes tonos involucrados en una señal, es ineficiente para describir sus respectivas ocurrencias en el tiempo. Sin embargo, a partir de este análisis se ha desarrollado un método dedicado al plano TF: la STFT, el cual es el método más utilizado para estudiar señales no estacionarias [34] y generaliza la transformada de Gabor al admitir una función ventana general [35]. Al deslizar una ventana a lo largo del eje

temporal y obtener un análisis espectral para cada intervalo donde se posiciona la ventana, se puede obtener una distribución TF [39].

Para estudiar las propiedades de una señal en el tiempo *t* se hace énfasis en la señal en ese momento y se suprime la señal en el resto de los instantes temporales, para lograrlo se requiere la multiplicación de la señal $x(t)$ por una función ventana $h\ (t)$ centrada en el tiempo *t*, para producir la señal de la **Ec. 3**:

$$x_t(\tau) = x(\tau)\ h(\tau - t) \tag{3}$$

La señal ventaneada $X_t(\tau)$ es una función de dos tiempos, el tiempo en el que estamos centrados, *t*, y el tiempo en marcha, τ . La función ventana se elige buscando no alterar la señal en el tiempo *t*, sino que se suprima el resto de la señal en los tiempos distantes del tiempo de interés [34]. Debido a estas propiedades de la señal, la transformada de Fourier de la señal ventaneada refleja la distribución de la frecuencia alrededor de ese tiempo, como se muestra en las **Ecs. 4** y **5**:

$$X_t(\omega) = \frac{1}{\sqrt{2\pi}} \int e^{-j\omega\tau}\ x_t(\tau)\ d\tau \tag{4}$$

$$= \frac{1}{\sqrt{2\pi}} \int e^{-j\omega\tau}\ x(\tau)\ h(\tau - t)\ d\tau \tag{5}$$

El módulo al cuadrado de $X_t(\omega)$, $|\mathrm{STFT}|^2$, es llamado espectrograma o respirosonograma en el campo de los SR, y ofrece una descripción de la energía de la señal en el plano TF. Por lo que, el espectro de densidad de energía de una señal $x(t)$ en el tiempo t y frecuencia ω está dada por la **Ec. 6**:

$$SP(t,\omega) = |X_t(\omega)|^2 = \left| \frac{1}{\sqrt{2\pi}} \int e^{-j\omega\tau} x(\tau) h(\tau - t) d\tau \right|^2 \tag{6}$$

donde $h(\tau - t)$ es una función ventana centrada en el tiempo t. Al multiplicarse por la función ventana $h(t)$, la señal original es enfatizada en el tiempo t y suprimida en el tiempo τ. Para cada instante de tiempo, se obtiene un espectro diferente, y la totalidad de este espectro es la TFR llamada SP.

Una de las ventajas del SP es su habilidad para reintroducir la noción del tiempo en el análisis de la señal, de manera que se observan los cambios instantáneos de la frecuencia. Representa, en un plano bidimensional, cómo el nivel de energía de la señal varía en el tiempo y la frecuencia [26], es una forma para visualizar la potencia, o intensidad, de una señal a lo largo del tiempo a diferentes frecuencias [32]. Sin embargo, esta técnica requiere una selección de parámetros que impactan directamente en la resolución en frecuencia de la TFR, como el tipo y la longitud de ventana, así como el traslape entre segmentos adyacentes.

Una de las limitaciones fundamentales en el método del SP es la elección de la ventana misma, ya que afecta el comportamiento de la representación al existir un límite inferior en la resolución conjunta TF. Si se restringe la ventana en el dominio temporal para obtener una mejor resolución en el dominio del tiempo se obtiene como resultado una ventana proporcionalmente más amplia con una localización más imprecisa y difusa en el dominio espectral [39]. Esta restricción, que es una derivación del principio de incertidumbre de Heisenberg, impulsó la búsqueda de otro tipo de transformadas de dominio mixto, como la transformada *wavelet* [35].

III.3.3. Espectro de Hilbert-Huang

Históricamente, el análisis de Fourier ha sido reconocido como el método general para examinar la distribución TF global de una señal, por lo que el término “espectro” se ha transformado casi en sinónimo de la transformada de Fourier. Este análisis ha sido aplicado a todo tipo de datos debido en parte a su relativa simplicidad y, a pesar de que es válido bajo condiciones generales, existen algunas restricciones a considerar al aplicar esta transformada, e.g., el sistema que genera la señal debe ser lineal y los datos deben ser estrictamente estacionarios. El no cumplir con estas condiciones resultará en un espectro que no ofrece

información importante sobre la señal, que no hará sentido para comprender los componentes de esta o que lleva a resultados erróneos o confusos [39].

El HHS es un método de análisis de datos que considera la energía regional en el plano local TF y se basa en el algoritmo EMD para obtener un conjunto de funciones IMF, las cuales son consideradas también como una expansión de los datos. La descomposición se basa en la extracción directa de energía asociada a diversas escalas de tiempo intrínsecas a la señal. Los IMF presentan transformadas de Hilbert bien definidas y es a partir de ellas que se puede calcular la frecuencia instantánea, además de poder localizar cualquier evento tanto en el eje de tiempo como en el eje de frecuencia [34]. Algunas características importantes de los IMF son: 1) estas funciones son derivadas de y basadas en los datos, por lo que el proceso es adaptativo, si los datos provienen de un sistema lineal la expansión será lineal, si provienen de un sistema no lineal la expansión también lo será, y 2) poseen información sobre la energía local y la frecuencia instantánea de la señal, datos que permiten conocer la distribución completa tiempo-frecuencia de la señal [39].

El método HHS comprende dos pasos para analizar los datos. La primera etapa consiste en realizar un preprocesamiento de datos aplicando un método de EMD, para descomponer la señal en sus componentes IMF [39]. Más adelante, se mencionarán algunos de estos métodos de descomposición, así como algunas ventajas y desventajas sobre los demás. La segunda etapa consiste en aplicar la transformada de Hilbert a las funciones IMF para con ellas construir la distribución tiempo-frecuencia, o HHS. Como se mencionó anteriormente, esta TFR requiere de los valores instantáneos de energía y frecuencia, en contraste con otras técnicas como el análisis de Fourier, que utilizan la frecuencia y la energía globales de la señal.

La frecuencia instantánea puede ser definida a partir de la señal analítica. En la **Ec. 7** se muestra una serie de tiempo arbitrario $x(t)$ [39], para la cual se calcula su transformada de Hilbert $y(t)$ como

$$y(t) = \frac{1}{\pi} P \left\{ \int_{-\infty}^{\infty} \frac{x(t')}{t-t'} \, dt' \right\} \tag{7}$$

donde $P\{\cdot\}$ indica el valor principal de Cauchy, el cual es un método que permite asignar valores a algunas integrales impropias que de lo contrario resultarían indefinidas. Mediante esta ecuación se define la transformada de Hilbert, haciendo énfasis en las propiedades locales de $x(t)$, como la convolución de $x(t)$ con $1/t$. De acuerdo con esta definición, $x(t)$ y $y(t)$ forman el par conjugado complejo, a partir de las cuales se puede obtener la señal analítica $z(t)$, como

$$z(t) = x(t) + iy(t) = a(t)e^{i\theta(t)} \tag{8}$$

tal que

$$a(t) = [x^2(t) + y^2(t)]^{\frac{1}{2}}, \qquad \theta(t) = \arctan\left(\frac{y(t)}{x(t)}\right) \tag{9}$$

donde $a(t)$ denota la amplitud instantánea, mientras que la fase $\theta(t)$ permite obtener la frecuencia instantánea.

La transformada de Hilbert proporciona una manera única de definir la parte imaginaria tal que el resultado sea la función analítica. En la **Ec. 8** la expresión de coordenadas polares expresa la naturaleza local de la representación, ya que es el ajuste local más adecuado de una función trigonométrica variable en fase y en amplitud para la señal $x(t)$. A pesar de la transformada de Hilbert, aún existe controversia al definir la frecuencia instantánea como

$$\omega = \frac{d\theta(t)}{dt}. \tag{10}$$

Esto dio origen a la introducción del término de "función monocomponente", realizada por Cohen [34], donde para cada instante de tiempo existe un solo valor de frecuencia pudiendo representar solamente un componente, de ahí que reciba ese nombre. Sin embargo, no se acuñó una definición exacta para saber si una función entra en ese

concepto. Por lo tanto, se prefirió adoptar el término "banda estrecha" o *narrow band*, en un intento de encontrar sentido a la frecuencia instantánea. El ancho de banda puede ser definido en términos de momentos espectrales de tal manera que para una señal de banda estrecha el número de extremos y cruces por cero sean iguales [39]. Las **Ecs. 7** a **10** muestran que la frecuencia definida por medio de la aproximación de fase estacionaria coincide con el mejor ajuste local de una función sinusoidal, por lo que no es estrictamente necesario contar con un periodo oscilatorio completo para poder definir un valor de frecuencia. De esta manera, se puede definir el valor de frecuencia instantánea en cada punto de una señal variable en el tiempo, incluso para una función que contenga un solo tono.

Cabe mencionar que, el HHS presenta ventajas al ser comparado con técnicas tradicionales como el análisis de Fourier, ya que el HHS es una técnica adaptable y permite el ajuste del método directamente a partir de los datos [39], así como que considera la energía regional en el plano local TF y proporciona una representación avanzada de las características de TF de señales multicomponentes. En el campo de los SR, el HHS ha sido utilizado para la detección de sibilancias [41]. Sin embargo, este método tiene un problema: debido a su naturaleza local, se pueden producir oscilaciones con escalas muy similares en distintos modos, lo que se denomina mezcla de modos (*mode mixing*, MM), el cual se abordará en la sección correspondiente a EMD.

III.3.3.1. Modos de oscilación intrínsecos

Existen condiciones necesarias para poder definir la frecuencia instantánea de un componente, como que las funciones sean simétricas con respecto a la media local cero, y tenga el mismo número de cruces por cero y extremos. Por lo que, la definición formal de estas funciones es que un IMF es una función que satisface dos condiciones: 1) en el conjunto de datos, el número de extremos y el número de cruces por cero debe ser igual o diferir como mucho por uno, y 2) en cualquier punto, el valor medio de la envolvente definida por el mínimo local y la envolvente definida por el máximo local es cero. La primera condición es similar a los requerimientos de la banda estrecha descritos anteriormente, mientras que la segunda condición es una idea novedosa, implementada por Huang et al. [39], la cual propone

modificar el requisito global clásico por uno local, y es necesaria para evitar fluctuaciones en la frecuencia instantánea provocadas por formas de onda no simétricas. Idealmente, sería necesario obtener la media local en una escala local de tiempo, sin embargo, definir este periodo de tiempo no es posible, por lo que se opta por emplear el mínimo y máximo locales para forzar una simetría local, y de esta manera mantener la adaptabilidad del método.

El nombre de los IMF fue seleccionado porque representa el modo de oscilación presente en los datos. Siguiendo este principio, el IMF de cada ciclo, definido por el número de cruces por cero, contiene un solo modo oscilatorio sin formas de onda complejas. De acuerdo con esta definición un IMF no está restringido a una señal de banda estrecha y tiene la posibilidad de ser modulado por amplitud o por frecuencia, además de que le brinda la flexibilidad de poder ser no estacionaria. Siguiendo esa definición de IMF se observa que la definición dada por la **Ec. 10** ofrece la mejor frecuencia instantánea, así como que un IMF al que se le aplicó la transformada de Hilbert puede expresarse por medio de la **Ec. 8**. Como se mencionó anteriormente, no es necesario contar con un periodo oscilatorio completo para poder definir la frecuencia instantánea, ésta puede ser definida en cualquier punto y con su valor cambiante de un punto a otro [39].

III.3.3.2. Descomposición empírica de modos

En algún instante de tiempo, los datos pueden contener más de un modo oscilatorio y la transformada de Hilbert sencilla no puede proveer una descripción completa del contenido en frecuencia para el conjunto completo de datos, por lo que la señal se debe descomponer en elementos más pequeños como los IMF. Huang et al. [39] propusieron un método nuevo para tratar las señales no estacionarias y no lineales realizando su descomposición, este método es directo, intuitivo y adaptable ya que se basa en los datos.

La descomposición se basa en tres suposiciones: 1) la señal tiene al menos dos extremos, un máximo y un mínimo, 2) la escala de tiempo está definida por el lapso entre los extremos, y 3) si los datos no contienen extremos, pero contienen puntos de inflexión, pueden ser derivados una o más veces para revelar los extremos. Los resultados finales pueden ser obtenidos fácilmente mediante la integración de los componentes. La esencia del método es

identificar empíricamente los modos de oscilación intrínsecos de acuerdo con sus escalas de tiempo características, y descomponer los datos de acuerdo con ellas [39].

El método consiste en una extracción sistemática, designada como proceso de cernido, donde es necesario que se cumplan ciertas condiciones en los datos para obtener un resultado que se pueda denominar IMF. Este algoritmo emplea la envolvente de la señal definida por los máximos locales y la envolvente definida por los mínimos locales, de forma separada. Una vez que los extremos se identifican, se emplea una función de spline cúbico para conectar los máximos locales y obtener la envolvente superior, y el mismo proceso se realiza empleando los mínimos locales para producir la envolvente inferior [39]. Ambas envolventes deben cubrir todos los datos que existen entre sí. Posteriormente, mediante la **Ec. 11**, se calcula su promedio (designado como m_1) y la diferencia entre la señal $x(t)$ y m_1, la cual corresponde al primer componente h_1, tal que

$$h_1 = x(t) - m_1 . \tag{11}$$

De acuerdo con las reglas descritas anteriormente, el término h_1 ha sido construido para cumplir con los requerimientos de un IMF, por lo que debería ser considerado como tal. Sin embargo, en las señales reales es común que se presenten excedentes o retrasos que generen nuevos extremos y desplacen a los extremos originales. Estos datos no tienen efecto directo sobre el resultado, ya que la señal de interés es la media, no las envolventes, y es aquella la que entra al proceso de cernido. Este algoritmo alcanza el objetivo de hacer más simétricos los perfiles de las ondas, y para lograrlo el procedimiento de cernido debe repetirse varias veces. En la siguiente etapa de cernido, h_1 será tratado como la señal o los datos originales, siguiendo

$$h_1 - m_{11} = h_{11} \tag{12}$$

tal que m_{11} sea considerado como el promedio calculado a partir de la señal h_1 y h_{11} sea el nuevo residuo. Este procedimiento se puede repetir k veces, hasta que el residuo h_{1k} cumpla con las características de un IMF, de forma

$$h_{1(k-1)} - m_{1k} = h_{1k} \tag{13}$$

Este resultado, con $h_{1(k-1)}$ como la nueva señal, m_{1k} como el nuevo promedio calculado y h_{1k} como el nuevo residuo, es entonces designado como

$$c_1 = h_{1k}\ , \tag{14}$$

el cual es considerado oficialmente como el primer IMF de la señal. Generalmente, c_1 contiene el componente de periodo más corto de la señal o su componente más fino. A partir de este resultado, c_1 puede ser separado del resto de datos mediante la **Ec. 15** para obtener el residuo r_1, siendo

$$x(t) - c_1 = r_1. \tag{15}$$

Ya que el residuo r_1 aún contiene información valiosa de componentes de periodos más largos, se emplea como la nueva señal y es sometido al proceso descrito anteriormente. Este proceso de cernido puede ser detenido cuando el residuo, r_n, llega a ser una función monotónica de la cual ya no se puede extraer ningún otro IMF. Como resultado de la descomposición, una señal $x(t)$ se puede expresar como una combinación lineal de sus componentes como:

$$x(t) = \sum_{i=1}^{n} c_i + r_n \quad (16)$$

donde n es el número de IMFs extraídos, c_i son los IMFs y r_n es el residuo de la señal original $x(t)$. Después de obtener los componentes IMF, la transformada de Hilbert se aplica a cada uno de ellos obteniendo una señal compleja, denominada señal analítica, a partir de la cual es posible estimar su frecuencia instantánea mediante la **Ec. 10** como función del tiempo. Este método permite obtener una representación de frecuencia y amplitud variables, de alta resolución TF conjunta, en la que algunos de sus componentes en frecuencia pueden ser no estacionarios mientras que otros permanecen siendo estacionarios [39].

Sin embargo, el método EMD tiene un problema: debido a su naturaleza local, se pueden producir oscilaciones con escalas muy similares en distintos modos, lo que deriva en el problema de la mezcla de modos. En el análisis de SR que contienen sonidos adventicios continuos se ha reportado que, debido al problema de MM, algunos componentes de frecuencias de las sibilancias aparecen en distintos IMF a la vez, lo que conduce a una separación pobre de escalas de frecuencia [42].

Para reducir este problema se han propuesto métodos como la EMD de conjunto (*ensemble empirical mode decomposition*, EEMD), la cual realiza la descomposición sobre un conjunto de copias de la señal original a las cuales se les agrega ruido blanco Gaussiano, y cuyo resultado se obtiene realizando el promedio [43]. Como resultado, se obtienen IMF más regulares, con escalas similares para todo el tiempo.

El algoritmo EEMD puede describirse en tres pasos: 1) generar n señales que contengan la suma de la señal original más diferentes realizaciones de ruido blanco de varianza unitaria media cero, 2) realizar una descomposición mediante EMD de cada una de las n señales obtenidas en el paso 1, para obtener sus IMF, y 3) promediar los modos calculados en el paso 2 para obtener un modo final. Sus parámetros de análisis son la relación señal-a-ruido (*signal-to-noise ratio*, SNR) para el ruido añadido, y el número de iteraciones a realizar, el cual al ser incrementado ayuda a reducir el nivel de ruido residual de los IMF obtenidos.

Algunas características notables de la EEMD son que la extracción de cada IMF requiere de un número distinto de iteraciones, y que cada señal originada en el paso 1 del algoritmo anteriormente descrito, es descompuesta de forma independiente a las otras realizaciones, donde para cada una de ellas se obtiene un residuo en cada etapa sin que haya una conexión entre las diferentes señales. Esto es causa de algunas desventajas del método, e.g., que las distintas realizaciones de la señal más ruido pueden producir un número diferente de modos, lo que dificulta el promedio final, que la descomposición no sea completa y que la señal reconstruida, i.e., la suma de los IMF y el residuo o tendencia final, contengan ruido residual [44].

El método de EMD de ensamble completo con ruido adaptativo (*complete ensemble empirical mode decomposition with adaptive noise*, CEEMDAN), desarrollado por Torres et. al. [45], ha resuelto este problema, además de reducir el error de reconstrucción hasta un nivel no significativo. Este método usa cada modo o IMF final para calcular el siguiente. Cada modo es calculado de forma secuencial, se estiman las medias locales de cada iteración que contiene señal más ruido y se define el IMF verdadero como la diferencia entre el residuo actual y el promedio de sus medias locales.

En este método la idea general es obtener los IMF a partir de una señal, donde el primer modo es calculado como en EEMD. Posteriormente, se obtiene un único primer residuo, independientemente del nivel de ruido elegido. Después, el primer modo de descomposición es calculado para un conjunto del primer residuo más diferentes realizaciones de un nivel particular de ruido. El segundo IMF es definido como el promedio de esos modos, y el siguiente residuo es calculado de la misma manera, al primer residuo se resta el segundo modo, tal como en el método de EEMD. Este procedimiento es repetido hasta alcanzar un criterio de paro, i.e., hasta obtener un residuo que no pueda ser descompuesto por EMD, debido a que no satisface las condiciones de un IMF o porque contenga menos de tres extremos locales.

De acuerdo con este método, una señal de interés $x(t)$ puede expresarse de acuerdo con la **Ec. 16**, esta ecuación asegura la propiedad del método de ser completo y proporcionar una capacidad de reconstrucción total y exacta de los datos originales [44]. El número final de modos se determina por los datos y el criterio de paro.

III.3.4. Transformada Synchrosqueezing

Otro de los métodos TF de alta resolución es la SST, la cual se ha aplicado con resultados prometedores para extraer información sobre la respiración a partir de otras señales fisiológicas, e.g., de fotopletismografía [46] o de ECG [47]. Cabe mencionar que, hasta nuestro conocimiento, en el campo de los SR la SST no se ha empleado para la clasificación automática de sibilancias.

El objetivo de emplear esta técnica TF es analizar señales multicomponente, definidas como una superposición de componentes modulados en frecuencia y componentes modulados en amplitud, definida como $x(t)$ de modo que

$$x(t) = \sum_{k=1}^{K} x_k(t) \quad \text{con} \quad x_k(t) = a_k(t) e^{2\pi j \theta_k(t)} \tag{17}$$

para una cantidad finita de k componentes tal que $K \in N$, $a_k(t)$ siendo la amplitud instantánea y $\theta_k'(t)$, la derivada de la fase, denota la frecuencia instantánea, donde x_k satisface $a_k(t) > 0$, $\theta_k'(t) > 0$ y $\theta_{k+1}'(t) > \theta(t)$ para todo instante de tiempo t. De modo que esta señal cuente con una TFR ideal ($TFRi$) definida como:

$$TFRi_x(t, \omega) = \sum_{k=1}^{K} a_k(t)\, \delta\big(\omega - \theta_k'(t)\big) \tag{18}$$

donde δ denota la delta de Dirac [48].

Uno de los objetivos de la SST es separar y recuperar los distintos modos que componen una señal multicomponente. Esta transformada se basa en una estimación de la frecuencia instantánea de cada modo, la cual es empleada para agudizar la energía de la representación alrededor de las crestas, para aproximar las curvas de tiempo y frecuencia instantánea $\big(t, \theta_k'(t)\big)$. Además, la SST tiene como meta principal emplear un proceso de reasignación de tal forma que la transformada resultante pueda ser invertible [48]. La SST

basada en la STFT se fundamenta en la definición del estimado de frecuencia instantánea local $\widehat{\omega}_x$:

$$\widehat{\omega}_x(t,\xi) = \frac{\partial \arg V_x^h(t,\xi)}{\partial \mathrm{t}} = \Re\left\{\frac{1}{2\pi j}\frac{\partial_t V_x^h(t,\xi)}{V_x^h(t,\xi)}\right\}, \qquad \text{donde sea que } V_x^h(t,\xi) \neq 0 \qquad (19)$$

donde $arg\ V_x^h(t,\xi)$ denota el argumento de la STFT de la señal $x(t)$ empleando una ventana h, mientras que ξ representa la transformada de la fase al estimar la frecuencia instantánea, dentro del conjunto de los reales $\Re$. La SST basada en STFT de $x(t)$ con umbral γ es obtenido al mover cualquier coeficiente $V_x^h(t,\xi)$ con magnitud mayor que γ a la ubicación $(t,\ \widehat{\omega}_x(t,\xi))$, de forma que:

$$SST_x^{\gamma}(t,\omega) = \int_{|V_x^h(t,\xi)|>\gamma} V_x^h(t,\xi)\ \delta(\omega - \widehat{\omega}_x(t,\xi))\mathrm{d}\xi\,. \qquad (20)$$

Cualquier modo $x_k(t)$ puede ser reconstruido al sumar los coeficientes de la SST basada en STFT alrededor de la k-ésima cresta, seleccionando únicamente los coeficientes relacionados con el k-ésimo modo, de forma que

$$x_k(t) \approx \frac{1}{h(0)} \int_{|\omega-\varphi_k(t)|<d} SST_x^{\gamma}(t,\omega)\ d\omega \qquad (21)$$

donde φ_k es la estimación de la frecuencia instantánea θ_k', y el parámetro d es empleado para compensar errores en la estimación de la frecuencia instantánea; idealmente se debe mantener lo suficientemente pequeño para evitar la MM.

Algunas de las ventajas de la SST basada en STFT es que la nitidez que ofrece esta representación se mantiene igual para frecuencias bajas, como para frecuencias altas, al contrario de otros métodos como la SST basada en transformada *wavelet* continua

(*continuous wavelet transform*, CWT) la cual ofrece mayor resolución a frecuencias bajas [48]. Además, la SST basada en STFT es un método adaptable que consiste en una técnica especial de reasignación, es robusto a diferentes tipos de ruido y debido a su naturaleza local logra detectar componentes que no existen a lo largo de todo el tiempo y por lo tanto se obtiene el comportamiento dinámico de la señal, además de que su representación proporciona información visual relevante. Sin embargo, la resolución en frecuencia impacta directamente en la nitidez de la representación, ya que la SST depende de dos parámetros principales: la elección del tipo y la longitud de la ventana que define la TFR [47].

III.4. Inteligencia artificial

En 1955, John McCarthy, uno de los pioneros de la inteligencia artificial (*artificial intelligence*, AI), fue el primero en acuñar el término y definirlo a grandes rasgos como que "la meta de la AI es desarrollar máquinas que se comporten como si fueran inteligentes". Sin embargo, esta definición no es suficiente, ya que la AI tiene como meta resolver problemas que son difíciles para muchos circuitos, sensores, robots, vehículos, entre otros dispositivos, problemas que están relacionados con una alta capacidad intelectual humana. Por lo que, años más tarde Elaine Rich volvió a definir a la AI como "el estudio de cómo hacer que las computadoras hagan cosas en las cuales, por el momento, las personas son mejores" [49]. Esta definición engloba las actividades que han realizado los investigadores en las últimas décadas, y que seguramente seguirá actualizado en los próximos años.

Existen actividades donde las computadoras superan las capacidades humanas, como la multiplicación de números grandes, la ejecución de múltiples cómputos u operaciones en un periodo corto de tiempo. Sin embargo, existen muchas más actividades superiores a las máquinas, pertenecientes a los humanos, que sirven de referencia para los programas computacionales de AI, como el seguimiento de patrones, la memorización de textos largos, la habilidad de reconocer el entorno, objetos o situaciones nuevas en fracción de segundos, y de ser necesario, tomar decisiones y planear acciones con base en la información.

Una de las cualidades humanas más importantes es la capacidad de adaptación, donde se emplea el aprendizaje, a través del conocimiento o reconocimiento del medio y condiciones ambientales para cambiar nuestro comportamiento [49]. Esta habilidad superior se ha tratado de transmitir a los sistemas computacionales, para desarrollar el aprendizaje de máquina o *machine learning*, el cual es una de las principales áreas de la AI, junto con el aprendizaje profundo. La diferencia entre la programación clásica y la AI es que, en el primero los humanos dictan las reglas al escribir un programa y los datos se procesan siguiendo dichas reglas o pasos para producir una respuesta. Mientras que, en el segundo, i.e., AI y *machine learning*, el usuario humano emplea datos y sus respuestas esperadas como entrada a un algoritmo y a la salida obtiene las reglas del procedimiento. Además, estas reglas pueden ser aplicadas a datos nuevos para obtener respuestas originales [50], por lo que se puede decir que un sistema de aprendizaje de máquina es entrenado.

En décadas recientes, inició una era donde físicos y científicos en computación y en ciencias cognitivas, lograron mostrar que las computadoras contaban con la capacidad y la suficiente potencia para emplear redes neuronales modeladas matemáticamente para realizar entrenamiento mediante ejemplos, que les permitieron aprender a reconocer patrones, e.g., símbolos escritos a mano, reconocimiento de rostros en fotografías, o el sistema Nettalk, capaz de aprender a hablar a partir de ejemplos de texto [49]. A pesar de las grandes capacidades de las redes, combinar las fórmulas empleadas para su desarrollo con el conocimiento humano presenta varias dificultades. Para sobrellevar este obstáculo diversas alternativas han sido propuestas, como el razonamiento probabilístico que emplea redes Bayesianas, lógica difusa, técnicas de aprendizaje de máquina como árboles de decisión, entre otros métodos.

Cerca del año 2010, después de casi tres décadas de investigación sobre redes neuronales, los científicos de esta área comenzaron a ver los frutos de su trabajo. Mediante la poderosa herramienta de redes de aprendizaje profundo han logrado clasificar imágenes con muy alta precisión, lo que ha llevado a toda clase de aplicaciones como vehículos autónomos, robots de servicio y ayuda en diagnóstico médico.

III.4.1. Redes neuronales y aprendizaje profundo

En el cerebro humano existen redes de células nerviosas conocidas como redes neuronales, las cuales están compuestas por cerca de 100 billones de células. Estas estructuras, complejas y adaptables, son la razón para que los humanos podamos aprender habilidades motoras e intelectuales. Durante varios siglos se han realizado esfuerzos por entender el funcionamiento del cerebro. A principios del siglo XX se llegó a la conclusión de que la estructura fundamental del cerebro, las neuronas, las células nerviosas y sus conexiones, son las responsables de estas funciones cruciales para la vida y el aprendizaje. Trasladar ese mecanismo de funcionamiento al campo de la AI es bastante complejo, ya que las herramientas existentes para realizar modelos, como las computadoras, sus lenguajes de programación o las matemáticas, son muy distintos al cerebro humano. Sin embargo, emplear redes neuronales artificiales ha brindado una nueva perspectiva. Al tomar como punto de partida el conocimiento sobre el funcionamiento de las redes neuronales naturales, estas se intentan modelar en equipos o hardware para realizar simulación o incluso su reconstrucción [49].

Al modelar matemáticamente una conexión entre neuronas, se calcula la "carga" de potencial de activación, la cual resulta de sumar valores de salida ponderados de las conexiones existentes con otras neuronas. Posteriormente, al resultado se aplica una función de activación, el cual pasa a las neuronas vecinas para repetir el procedimiento. Dentro del conocimiento fisiológico es bien sabido que una de las formas de aprendizaje consiste en reforzar conexiones sinápticas al incrementar los impulsos eléctricos que se transmiten a través de ciertos enlaces. De manera que este principio se transfiere a las redes sintéticas como que, cuando hay una conexión entre la neurona *a* y la neurona *b*, y numerosas señales son enviadas de la neurona *a* a la neurona *b*, esto resulta en la actividad simultánea de ambas neuronas, reforzando el peso entre esa conexión particular [49]. Además, el cambio en esos pesos puede depender de una constante tal que determine el tamaño de los pasos individuales de aprendizaje de la red, llamada tasa de aprendizaje.

Al utilizar redes neuronales se debe considerar usar representaciones significativas de entrada para aprender las características deseadas, en un intento de que la red conduzca a las

salidas deseadas. Al elegir una representación o transformación de datos de entrada apropiada los datos se hacen más adaptables a la tarea de la red en cuestión. Una representación puede considerarse como una forma diferente de visualizar los datos, e.g., los formatos de color de una imagen como RGB (rojo-verde-azul, *red-green-blue*), HSV (valor de saturación de tono, *hue-saturation-value*), o las diferentes TFR de una misma señal [50].

Generalmente, estas redes están compuestas por una etapa de preprocesamiento, que puede variar dependiendo de la aplicación, pero por lo general las variables de entrada pasan por un proceso de normalización. En el caso de fotografías, se suele transformar su tamaño para reducir su espacio de dimensiones. Posteriormente, la red comprende dos partes. La primera parte comprende capas pre-entrenadas por aprendizaje no supervisado, donde cada capa representa características del patrón de entrada, entre más baja sea la capa las características que extrae son más simples. Cuando se emplean fotografías como entrada a la red para realizar reconocimiento de objetos, los rasgos de las capas más bajas suelen representar líneas o bordes (horizontales, verticales, o en distintas orientaciones), mientras que las características más complejas, como un objeto completo o un rostro humano, se forman en las capas más altas. La segunda parte de la red comprende capas que emplean aprendizaje supervisado, las cuales pueden ser entrenadas por retropropagación, donde para cada ejemplo de entrenamiento la salida de la última capa de la parte 1, una capa de características es utilizada como entrada y una etiqueta será la salida objetivo. Finalmente, se requiere una forma de medir el desempeño del algoritmo para determinar el error entre la salida actual y la salida esperada. Esta medida es empleada como retroalimentación para ajustar el algoritmo buscando reducir el error obtenido [49]. La función de pérdida, o *loss function*, la cual expresa qué tan bien es realizada la tarea de la red, es la encargada de llevar a cabo esta medición [50].

Un ejemplo de una red neuronal se muestra en la **Fig. 7** [50], donde se muestra la estructura general de una red al transformar la imagen de un dígito en diversas representaciones que son visiblemente diferentes a la imagen original, las cuales ofrecen y recopilan diversa información sobre el resultado final. Se puede considerar a una red como una operación donde la entrada atraviesa filtros sucesivos para finalizar como información muy sencilla, y de esta manera ser clasificada en un grupo específico.

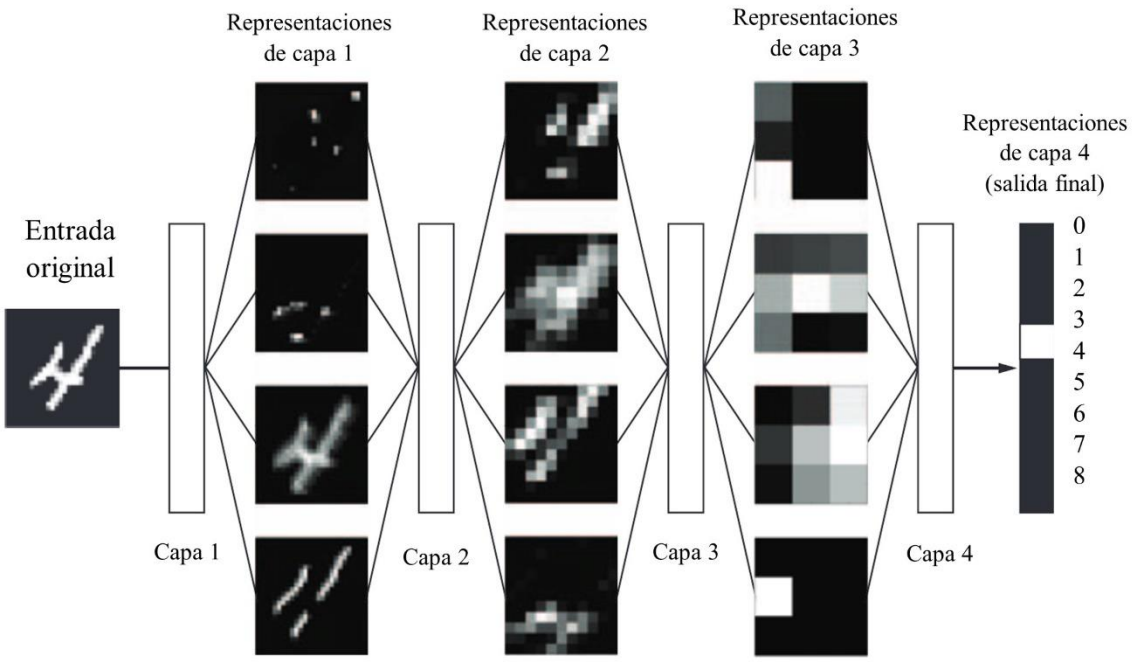

Fig. 7. Representaciones profundas aprendidas por un modelo de clasificación de dígitos [49].

Cada capa de una red realiza una operación sobre sus datos de entrada, la cual es parametrizada en sus pesos, también llamados parámetros. Por lo que la red, al aprender, encuentra un conjunto de valores para los pesos de las capas, de tal forma que los ejemplos de entrada se asocien correctamente a sus etiquetas de salida [50]. Es por esto que, el valor de los parámetros a utilizar en cada entrenamiento causa un gran impacto en el resultado final y en su calidad.

Una de las novedades que presentan las redes neuronales es que permite a un modelo aprender sobre una representación empleando todas sus capas a la vez, en lugar de aprender sucesivamente cada una de las capas. Al emplear este método conjunto, en cada momento en que el modelo ajusta uno de sus parámetros internos, el resto de las características que dependen de él se adaptan al cambio automáticamente, sin requerir de intervención o supervisión humana. Una aproximación para encontrar los hiperparámetros óptimos para un algoritmo de aprendizaje suele ser la implementación de búsqueda aleatoria, descenso de gradiente o una búsqueda exhaustiva que pruebe todas las diferentes combinaciones de

valores con el objetivo de minimizar el error al evaluar el conjunto de datos de validación [49].

Las redes neuronales requieren de un intensivo cálculo computacional, el cual puede incrementar debido al tamaño del vector de entrada, a un gran número de capas en la red y a la cantidad de datos necesarios para entrenarla. Sin embargo, existen diversas técnicas que pretenden reducir la complejidad y el tiempo empleado para entrenar estas redes, como alternar el tipo de capas que comprenden la red, e.g., emplear una capa de convolución, con un filtro lineal entrenable, y una capa de agrupación o *pooling*, empleando una función promedio o máximo calculada a partir de las neuronas de entrada. Así como utilizar capas de características que no estén completamente conectadas (*fully connected layer*), para que cada neurona de características retenga solo una porción de información de pocas neuronas provenientes de capas inferiores.

Una de las etapas más empleadas en las CNN es la etapa de convolución, la cual tiene por propósito crear conexiones locales de las capas anteriores y asignar esos datos o pesos a las capas siguientes, para emplearlos como un mapa de características [32]. La operación de convolución 2D se muestra en la **Ec. 22:**

$$y_i^n = \sum_i y_i^{n-1} * \omega_{ij}^n + b_i^n \tag{22}$$

donde y_i^{n-1} son los datos de entrada provenientes de la salida de las capas convolucionales anteriores, ω_{ij}^n es la *n*-ésima matriz de pesos, * es la operación de convolución y b_i^n es el *n*-ésimo vector de sesgo. Mientras que la unidad lineal rectificada, o función ReLU, es la función de activación más utilizada en las CNN, ya que previene problemas de gradiente dentro de la función de activación sigmoide [32]. La función de activación ReLU, puede definirse como:

$$r_i^n = \max(0,\ y_i^n) \tag{23}$$

donde r_i^n es la *i*-ésima salida de la *n*-ésima capa ReLU. Otra de las operaciones más empleadas es la agrupación, o *pooling*, encargada de reducir el número de muestras o datos mediante la reducción del tamaño de las matrices, ayudando a disminuir el costo computacional y a prevenir el sobreajuste. Las capas de agrupación más comunes son las que emplean el valor máximo, *max-pooling*, y el valor medio, *average-pooling* [32]. Estas capas de agrupación operan al extraer ventanas de los mapas de características y dar como resultado un único valor de esa matriz [50]. El cálculo general de esta operación se expresa en la **Ec. 24**:

$$p_i^n = \ operación\ máximo\ o\ promedio\ \{r_i^n\}\ . \tag{24}$$

En una capa completamente conectada, o capa *flatten*, se realiza la conjunción de todas las matrices provenientes de capas anteriores, las cuales se "aplanan" y se conectan a otras capas completamente conectadas [32]. Mientras que la función *cross-entropy* es empleada en la etapa de retropropagación, la cual se encarga de brindar información a capas anteriores sobre la distancia que existe entre la dispersión de los valores predichos por la red y la dispersión de las categorías verdaderas de los datos. El cálculo de esta función puede expresarse como:

$$H\ (y_i^t, y_i^n) = -\sum_{j=1}^{K} y_i^t \log y_i^n \tag{25}$$

donde y_i^t representa los valores verdaderos de los datos, mientras que y_i^n simboliza los valores predichos por la red.

Una de las ventajas que presentan las redes neuronales es su memoria asociativa, al poder reconocer patrones. Si un patrón o imagen se aprende durante el entrenamiento, el mismo patrón puede ser reconocido en el conjunto de validación a pesar de que éste se muestre en otro lugar de la imagen (traslación), cambie de tamaño, sea más grande o pequeño, o incluso si aparece rotado. Además de que las características de los datos son extraídas automáticamente por las redes, en comparación a algoritmos de *machine learning* donde los científicos de datos tenían el trabajo de encontrar ese conjunto de características manualmente [49], [50].

Al emplear una imagen como entrada a la red, nos encontramos con un problema importante: la gran dimensión de datos o pixeles que contiene. Si multiplicamos el gran tamaño de la imagen, vector de entrada, y el número de imágenes tendremos una red que necesita mucho tiempo de cómputo para ser entrenada. Por lo que una medida para mantener el tiempo de cálculo dentro de límites adecuados es reducir la dimensión de los datos de entrada [50]. Como se mencionó en el párrafo anterior, el mapeo o selección usualmente era creado de forma manual. Sin embargo, en muchas aplicaciones, como la clasificación de objetos en imágenes es complicado, si no que imposible, encontrar, de forma manual, una fórmula general para extraer las características importantes que se quieren buscar. En este aspecto, las redes neuronales han sobrellevado esta dificultad al automatizar este proceso y disminuir el tiempo de procesamiento de los datos. Esta medida se refuerza a partir de otra de las desventajas que presentan los modelos de entrenamiento, la cual requiere una cantidad grande de datos para obtener resultados precisos.

Diversas aplicaciones de redes neuronales incluyen el reconocimiento y clasificación de patrones, dígitos u objetos en imágenes, describir el contenido de una fotografía en una oración, o reconocer el contenido de texto escrito a mano o lenguaje hablado [50]. Aplicaciones más creativas han integrado la posibilidad de componer melodías, generar texto, crear imágenes con distintas características, estilos y contenido, o incluso corregir códigos de computadora. Además, se han aprovechado las características y áreas de oportunidad de este tipo de redes en el área industrial, aplicándolas en reconocimiento de voz y rostros, control de coches autónomos, o pronosticar precios de acciones de la bolsa.

III.4.2. Transferencia de aprendizaje

La transferencia de aprendizaje, o *transfer learning*, es un método empleado en aprendizaje profundo donde un modelo entrenado previamente para una tarea se emplea como punto de partida para aprender una nueva tarea que puede ser similar. Utilizar una red pre-entrenada resulta más rápido y sencillo que entrenar una red nueva desde el principio con pesos al azar inicializados de cero. Esta técnica puede transferir las características aprendidas previamente a una aplicación nueva empleando una cantidad menor de imágenes de entrenamiento y es usada en detección de objetos, reconocimiento de imágenes y de voz, clasificación, entre otras [51].

Algunas ventajas de *transfer learning* son que permite entrenar modelos a pesar de que se tengan bases de datos etiquetados de tamaño pequeño, al utilizar redes habituales que ya han sido entrenadas previamente con miles o millones de datos de muestra. Además, reduce el tiempo de entrenamiento y la memoria necesaria para los recursos informáticos, ya que los pesos inician en valores aprendidos con anterioridad en vez de aprenderse desde cero, y se aprovecha la estructura de modelos desarrollados por científicos de datos e investigadores de aprendizaje profundo [52].

Generalmente, los pasos de trabajo de este tipo de redes son [53]:

1) Seleccionar un modelo previamente entrenado con una base de datos grande.
2) Reemplazar las últimas capas, ya que la red fue entrenada con un conjunto de datos específicos, se deben reemplazar las últimas capas del modelo para ajustarlas al nuevo conjunto de datos y que su última capa completamente conectada contenga la misma cantidad de salidas o clases nuevas, además se deben considerar las funciones de clasificación ya que dependiendo de si es una salida binaria o multiclase se pueden elegir diversos tipos de capa.
3) Mantener pesos de capas anteriores de la red con tasas de aprendizaje en cero, este es un paso opcional que puede ahorrar tiempo en el entrenamiento, ya que al no actualizar los parámetros de algunas capas se acelera el proceso y evita que la red se sobreajuste si el conjunto de datos es pequeño.

4) Volver a entrenar el modelo, se refiere a realizar un entrenamiento con el conjunto de datos y categorías nuevas, el cual actualizará la red para que aprenda las nuevas características.
5) Hacer predicciones y evaluar el desempeño de la red, al obtener los resultados empleando el nuevo conjunto de imágenes se puede evaluar qué tan bien realizó la clasificación.

Una buena forma de enfocar el proyecto a la tarea nueva es probar varios modelos para encontrar el que mejor resultados ofrezca y mejor se adapte a la aplicación. Afortunadamente, en la actualidad es relativamente sencillo encontrar programas de computación que ofrezcan estas redes pre-entrenadas, las cuales varían en tamaño, cantidad de capas, número de parámetros entrenables, así como en el conjunto de imágenes usado para entrenar las redes [52], [53]. Además, estos programas facilitan el entrenamiento con bases de datos pequeñas al ofrecer herramientas para generar aumento de datos: como la reflexión, rotación y reescalado de imágenes, lo cual ayuda a evitar que la red se sobreajuste al memorizar las imágenes de entrenamiento.

Capítulo 4

Metodología

IV.1. Metodología general

Para realizar la comparación de las diversas TFR a explorar en este trabajo de tesis, se empleó una señal de sonido sintético, particularmente un sonido adventicio continuo polifónico con el propósito de simular una sibilancia [9], [43]. Contar con una señal creada artificialmente permite conocer su TFR ideal como una referencia para realizar una comparación con las TFR a utilizar en este trabajo. Posteriormente, los parámetros elegidos se aplicarán a sonidos provenientes de una base de datos pública, que contiene SR con y sin sibilancias.

En la **Fig. 8** se muestra el diagrama de bloques de la metodología a seguir para realizar este trabajo de tesis, que incluye: 1) la búsqueda de una base de datos de SR de acceso público y el preprocesamiento de las señales, 2) la simulación de una señal sintética con características de sibilancia y la obtención de su TFR ideal, 3) la exploración y el cómputo de las TFR a emplear con un conjunto previamente seleccionado de parámetros, 4) la selección de la TFR más representativa para cada técnica TF, 5) el uso de las representaciones obtenidas para emplearlas como entrada a una red neuronal previamente entrenada para realizar la clasificación de estas imágenes en SR normales o SR con sibilancias y 6) la evaluación del desempeño del detector.

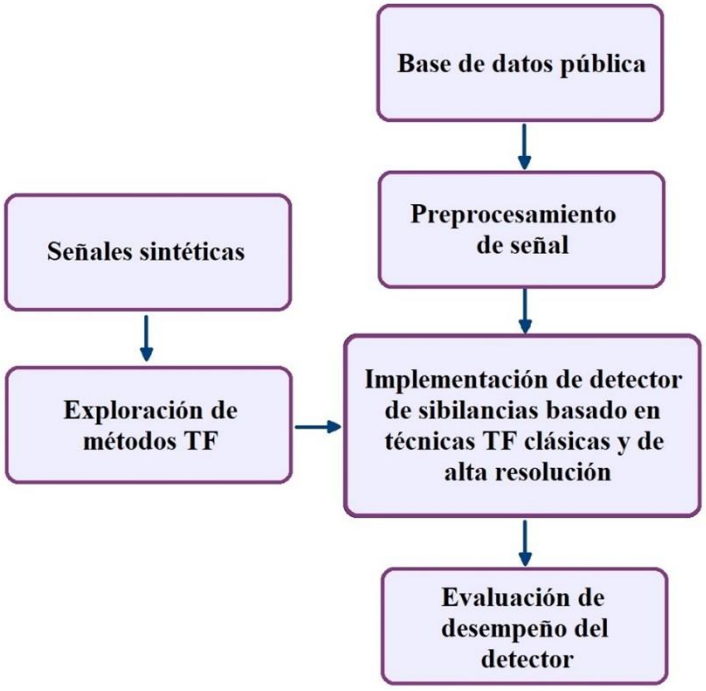

Fig. 8. Diagrama de bloques del método de detección.

IV.2. Simulación de señales sintéticas

Este trabajo se enfoca en la detección automática de ciclos respiratorios que contienen sibilancias, por lo que inicialmente se realizó la simulación de una señal que cumple con las características básicas temporales y espectrales de una sibilancia, i.e., que tenga una duración mínima de 100 ms y un componente principal de frecuencia mayor a 100 Hz. En esta sección se aborda la simulación computacional de una sibilancia sintética y su TFR ideal.

IV.2.1. Sibilancia simulada y su representación tiempo-frecuencia ideal

Un aspecto fundamental del algoritmo de detección de sibilancias es la técnica TF empleada, así como el ajuste de sus parámetros, por lo que se emplearon sibilancias simuladas que permitan ganar entendimiento de las técnicas propuestas para posteriormente comparar los métodos TF clásicos y de alta resolución seleccionados.

La sibilancia sintética fue simulada siguiendo el trabajo de Lozano et al. [43], mediante la conjunción de la **Ec. 26**, que representa la señal de un sonido adventicio continuo monofónico con frecuencia modulada sinusoidalmente dada por:

$$C_1(t) = sen[2\pi 80t + 0.6\, sen(2\pi 15t)] \qquad t \in [0, 0.3]\ s \qquad (26)$$

con la **Ec. 27**, que pertenece a la señal de un sonido adventicio continuo monofónico con frecuencia modulada linealmente:

$$C_2(t) = sen[2\pi 150t + 2\pi 150t^2] \qquad t \in [0.025, 0.275]\ \text{s.} \qquad (27)$$

Obteniendo una señal de sonido adventicio continuo polifónico, que resulta de la suma de ambos componentes modulados en la frecuencia, mediante la **Ec. 28**:

$$C_3(t) = \begin{cases} C_1(t) & t \in (0, 0.025] \\ C_1(t) + C_2(t) & t \in (0.025, 0.275] \\ C_1(t) & t \in (0.275, 0.3] \end{cases} \qquad (28)$$

La señal anterior es representativa, ya que, como se mencionó en la sección I.2.2, los SR adventicios continuos pueden presentar uno o más componentes sinusoidales o lineales de frecuencia, así como diversos tonos. Además, en la **Fig. 6** se muestra un ejemplo de cómo se visualizan las sibilancias en un SP, apareciendo como líneas cuasi-horizontales cuya frecuencia puede variar a lo largo del tiempo. Debido a que las sibilancias pueden ser multicomponentes, se seleccionó la **Ec. 28** para representar dichas características en una simulación, de modo que la señal sintética se acercara a un sonido adventicio continuo real.

IV.3. Selección de parámetros de las representaciones tiempo-frecuencia

La selección de los parámetros para cada TFR se realizará mediante las señales sintéticas que cumplen con las características temporales de una sibilancia, simulando sonidos adventicios continuos monofónicos con frecuencias moduladas sinusoidal y linealmente, así como sonidos adventicios continuos polifónicos [43]. A partir de las **Ecs. 26-28** es posible obtener su frecuencia instantánea, permitiendo generar su TFR ideal teórica.

El cómputo de una TFR mediante cada técnica depende de distintos parámetros como la longitud y tipo de ventana para el SP y la SST, así como la técnica de EMD y el número de modos a emplear para la construcción del espectro en el HHS. Cabe mencionar que, la técnica de WVD no cuenta con parámetros modificables para realizar la TFR de una señal. Como consecuencia, el desempeño de cada técnica TF dependerá de la selección de dichos parámetros, siendo más o menos parecida a la TFR ideal. Esta elección de parámetros resulta de vital importancia, ya que como se señaló en la sección de Espectrograma, la selección de una ventana demasiado corta ofrece buena resolución temporal pero la resolución en la frecuencia es baja y viceversa, por lo que el objetivo de esta sección será encontrar los valores que ofrezcan una TFR adecuada y cercana a la TFR ideal. En la **Tabla I** se muestran los parámetros empleados en la búsqueda de la representación más cercana a la ideal, para cada

TFR, donde el traslape de ventanas para los métodos que emplean tal función es de 50%. Posteriormente, se realizó una comparación cualitativa y cuantitativa, mediante el método de información mutua (*mutual information*, MI), entre las imágenes TF obtenidas al utilizar los distintos parámetros. Cabe mencionar que los valores de energía de todas las TFR fueron normalizados, al realizar una relación entre su valor mínimo y máximo para reasignar su distribución en el rango [0,1] o [0,255] hablando de una imagen.

Tabla I. Parámetros para búsqueda de TFR más cercana a la ideal.

TFR	SP	SST	HHS
Tipo de Ventana	Hamming, Han, Blackman, Kaiser	Hamming, Han, Blackman, Kaiser	NA
Longitud (ms)	15, 32, 43, 75, 100, 125, 150, 175, 200	15, 32, 43, 75, 100, 125, 150, 175, 200	NA
Tipo de descomposición	NA	NA	EMD, CEEMDAN

***NA =** No Aplica para la TFR

IV.4. Base de datos ICBHI

Debido a la situación generada por la pandemia COVID-19, no fue factible realizar la adquisición de datos reales y se empleó una base de datos de SR que contiene sibilancias y sonidos ventilatorios normales. En particular, se utilizó la base de datos de señales de SR perteneciente al ICBHI *Challenge* 2017 [21]. Esta base de datos contiene 920 registros de audio pertenecientes a 126 pacientes, donde cada registro contiene anotaciones realizadas por médicos, neumólogos y cardiólogos, expertos en reconocimiento visual y auditivo de crepitancias y sibilancias. Además, se incluyen archivos que indican la duración de cada ciclo respiratorio, i.e., el periodo que comprende la inspiración y la espiración, así como la etiqueta de clasificación de estos ciclos respiratorios en una de las siguientes cuatro categorías: 1) sonidos normales, 2) sibilancias, 3) crepitancias, o 4) ambos sonidos adventicios (crepitancias y sibilancias). La base de datos incluye un total de 6,898 ciclos respiratorios, de los cuales 886 contienen sibilancias y 3,642 son normales, como se indica en la **Tabla II**. Cabe mencionar que, la cantidad total de ciclos respiratorios con sibilancias y normales, mencionados anteriormente, están distribuidos en toda la base de datos, donde se encuentran ciclos con sibilancias en las categorías 2 y 4, así como ciclos respiratorios normales en los 4 grupos.

Esta base de datos contiene muestras de audio recolectadas independientemente por dos equipos de investigación, en Portugal y en Grecia. Para realizar las adquisiciones, fue obtenida la aprobación de los comités de ética de las instituciones correspondientes. El primer equipo de investigación, localizado en la Escuela de Ciencias de la Salud de la Universidad de Aveiro (ESSUA en Porto, Portugal) realizó los registros sobre la tráquea y seis lugares del pecho sobre el lado izquierdo y el derecho, adquiriendo SR desde el tórax en la parte anterior,

Tabla II. Ciclos respiratorios totales de base ICBHI Challenge 2017

Clasificación	No. Ciclos respiratorios
Normal	3,642
Con crepitancias	1,864
Con sibilancias	886
Con crepitancias + sibilancias	506
No. total de ciclos respiratorios	6,898

posterior y lateral; participaron sujetos de todas las edades y algunos padecían infecciones respiratorias del tracto superior e inferior, neumonía, EPOC, asma, bronquiolitis, bronquiectasia y fibrosis quística. Se utilizaron diversos equipos e instrumentación (secuencialmente con Estetoscopio digital Welch Allyn Meditron Master Elite Plus Stethoscope modelo 5079-400; simultáneamente con 3M Littman Classic II SE; simultáneamente con Micrófonos electret air-coupled C 417, AKG Acoustics).

El segundo equipo de investigación, localizado en la Universidad Aristotélica de Thessaloniki (AUTH) en el hospital general G. Papanikolaou y el hospital general de Imathia, Grecia, realizó los registros de SR secuencialmente en seis posiciones del pecho. Siguiendo un protocolo, registraron los SR de pacientes mientras se encontraban sentados, produjeron eventos de tos, habla, risa y aclaramiento de garganta, empleando distintos dispositivos de registro de SR (Estetoscopio digital Welch Allyn Meditron Master Elite Plus Stethoscope modelo 5079-400 o 3M Littman 3200). Los participantes eran adultos y ancianos, con padecimientos de EPOC y comorbilidades, e.g. insuficiencia cardiaca, diabetes, hipertensión, entre otras.

IV.4.1. Preprocesamiento de datos

La base de datos está compuesta de las señales de SR registradas, información clínica de los sujetos y archivos con las anotaciones realizadas por los profesionales de la salud, indicando el inicio y fin de cada ciclo respiratorio y el inicio y fin de los eventos anormales respiratorios, i.e., crepitancias, sibilancias. Por lo que se realizó la compilación de todos los datos en un solo documento de manera que se pudieran seleccionar solamente los archivos que contienen sibilancias y los archivos de SR normales. Esto permitió seleccionar una muestra aleatoria por grupos, considerando los registros que contienen sibilancias y los registros de señales normales, lo cual resultó fundamental para plantear y analizar el desempeño de algoritmos de clasificación de sibilancias basados en las características TF de los SR normales y con sibilancias.

La selección de datos para evaluar cada ciclo respiratorio se realizó debido a las características de la base de datos, ya que en ella se encuentran señales de SR y archivos de

texto que indican el inicio y término del ciclo respiratorio. Además, nos basamos en los trabajos encontrados en la literatura, donde realizan clasificación por ciclo respiratorio etiquetado por los expertos. Considerar emplear una segmentación por ventanas más cortas pudiera afectar a la duración de las sibilancias, ya que las podría cortar y de esta manera la red no reconocería la sibilancia completa. Emplear un procesamiento por ciclos respiratorios permite visualizar los sonidos adventicios completos, los cuales también presentan variación en la duración. Sin embargo, puede resultar útil emplear esa segmentación en una investigación futura.

Con el objetivo de contar con grupos de datos balanceados, tanto en número de ciclos respiratorios como en número de pacientes, se seleccionó la misma cantidad de ciclos con sibilancias y ciclos normales. La **Tabla III** muestra la selección por grupos realizada en la base de datos ICBHI, donde el primero cuenta con 592 ciclos respiratorios que fueron marcados por los expertos como ciclos que contienen sibilancias y se encuentran dentro de la categoría 2) sibilancias. En la **Tabla II** se observa que en toda la base de datos existen 886 ciclos con sibilancias, sin embargo, 294 de ellos están contenidos en registros de SR de sujetos que presentan ambos sonidos adventicios (crepitancias y sibilancias), por lo que no fueron considerados para agregarlos en este trabajo, ya que nos enfocamos en el grupo de sujetos que únicamente presentan SR con sibilancias.

Los 592 ciclos del grupo que contiene sibilancias están contenidos en 132 registros de audio, pertenecientes a 48 pacientes. Mientras que, para el grupo de sujetos normales se seleccionaron aleatoriamente el mismo número de ciclos respiratorios normales dentro de la categoría 1) sonidos normales; los cuales se encuentran en 80 registros de audio, pertenecientes a 47 pacientes distintos.

Las señales contenidas en la base de datos cuentan con una frecuencia de muestro de 4 kHz, 10 kHz o 44.1 kHz, por lo que se estableció una sola frecuencia de muestreo de 4 kHz para estandarizar el procesamiento posterior, cumpliendo además con el criterio de Nyquist para SR. El remuestreo fue realizado en el software MATLAB 2022b (The MathWorks, Inc., Natick, MA, USA), mediante la función *resample(), la cual* aplica un filtro pasa-baja anti-*aliasing* FIR a la señal de entrada y compensa el retardo introducido por el filtro. Además,

Tabla III. Detalles del grupo normal y grupo con sibilancias

Grupo	Número de registros de audio	Número de pacientes	Número de ciclos respiratorios
Con sibilancias	132	48	592
Normales	80	47	592

se realizó una aproximación fraccionaria racional empleando las frecuencias de muestreo de 10 o 44.1 kHz, dentro de la tolerancia por default, mediante la función *rat()*.

Además, se aplicó un filtro FIR pasa altas de fase cero, con frecuencia de corte de 100 Hz, seleccionada de acuerdo con las características de las sibilancias presentadas en la literatura [9], [19], así como para minimizar la presencia de sonidos cardiacos, ruidos musculares y ruido ambiental. Así mismo, se empleó la función de MATLAB *normalize()*, la cual normaliza los datos del vector *x* perteneciente a la señal de SR.

IV.5. Métodos de detección automática mediante aprendizaje profundo

Se planteó el uso de redes neuronales para realizar detección de ciclos respiratorios normales y con sibilancias, para aprovechar las bondades que provee, ya que permiten emplear representaciones significativas de una señal como una imagen, o en este caso una TFR, para aprender sus características principales de manera que la red conduzca a las salidas deseadas. Además, debido a la complejidad y el alto costo computacional, se empleó la técnica de transferencia de aprendizaje para reducir el tiempo empleado para entrenar la red, así como aprovechar la capacidad de detección de este método al emplear bases de datos relativamente pequeñas como entrada.

IV.5.1. Detector de ciclos respiratorios normales y con sibilancias

Para realizar la detección de ciclos respiratorios normales y ciclos respiratorios con sibilancias, se realizaron diversas aproximaciones en cuanto a la clasificación. El primer

método consistió en implementar seis distintas redes neuronales sencillas, que incluyen una o más de las siguientes capas y funciones: convolución 2D, capa completamente conectada, aplanamiento o *flatten*, ReLU, *max-pooling* 2D, *dropout*, sigmoide y *crossentropy*. Un ejemplo de arquitectura de estas redes sencillas fue conformado por las capas:

1) Entrada de imágenes
2) Convolución 2D
3) Función ReLU
4) *Max-pooling*
5) Convolución 2D
6) Función ReLU
7) *Max-pooling*
8) Convolución 2D
9) Función ReLU
10) *Max-pooling*
11) *Flatten*
12) *Dropout*
13) Capa completamente conectada
14) Sigmoide
15) *Crossentropy.*

Para realizar las pruebas preliminares se empleó un subconjunto de 50% de los 1150 datos totales a emplear en este trabajo de tesis, ya que se realizó en una etapa temprana donde no se tenían todos los datos procesados. Al realizar el entrenamiento con las redes sencillas se obtuvieron desempeños de 50%, como el mostrado en la **Fig. 30**, por lo que se procedió a explorar otras redes y métodos de clasificación. Como una siguiente prueba se reprodujo la implementación de CNN realizada por Demir et al. [32], cuya estructura de red cuenta con dos ramas paralelas que emplean funciones de *average-pooling* y *max-pooling* en un esfuerzo para incrementar el desempeño de clasificación. Sin embargo, los resultados obtenidos mediante esta CNN también fueron bajos, con una tasa de clasificación de 50%. Posteriormente, se procedió a explorar el método de transferencia de aprendizaje, ya que en

la literatura ha reportado buenos resultados en trabajos que no emplean métodos auxiliares de clasificación o postprocesamiento de datos [20].

Al aplicar el método de *transfer learning* se realizaron entrenamientos mediante algunas de las redes comúnmente empleadas en clasificación de imágenes y, particularmente, en la detección de sibilancias reportada en la literatura, usando las arquitecturas de las redes mostradas en la **Fig. 9**: *SqueezeNet*, *GoogleNet*, *AlexNet*, y *VGG-16*, empleando el software MATLAB 2022b.

Al realizar las pruebas preliminares con las redes anteriormente mencionadas y el subconjunto de la mitad de los datos, se obtuvieron resultados de clasificación de 50% para las redes *AlexNet* y *VGG-16*, mientras que para la red *SqueezeNet* se obtuvo un desempeño de 56.6% y para la red *GoogleNet* un resultado de 63.3%. De acuerdo con estos resultados, y por cuestión de tiempo, se optó por detener la exploración de arquitecturas y emplear la red *GoogleNet*, la cual fue pre-entrenada con el conjunto de imágenes *ImageNet*, seleccionándola como arquitectura base para realizar todos los entrenamientos posteriores.

En la **Fig. 10** se muestra la estructura de la red *GoogleNet*, la cual cuenta con una capa de entrada para normalizar el tamaño de las imágenes (bloque color azul), que consta de las imágenes correspondientes a las TFR obtenidas mediante las cuatro técnicas planteadas. Las imágenes de las TFR deben ser redimensionadas para alcanzar un tamaño adecuado para la red, de 224 x 224 pixeles, donde cada imagen corresponde a un ciclo respiratorio etiquetado. Así mismo, la primera etapa de la red contiene capas de convolución 2D (bloques amarillos), funciones ReLU (cuadros color naranja), capas de agrupamiento o discretización *max-pooling* (color morado) y funciones de normalización (bloques verdes), esta etapa inicial es empleada para reducir la dimensión de las imágenes. Posteriormente, la

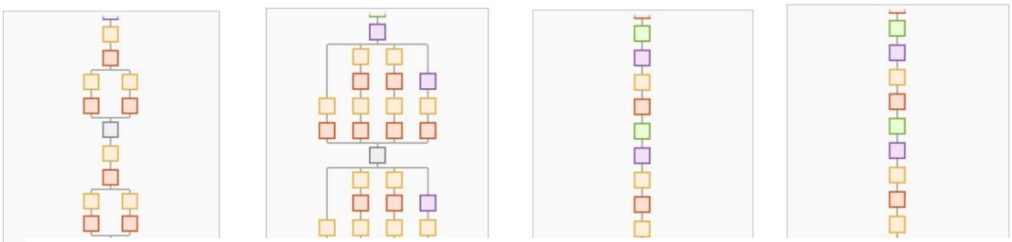

Fig. 9. Redes pre-entrenadas empleadas en transfer learning.

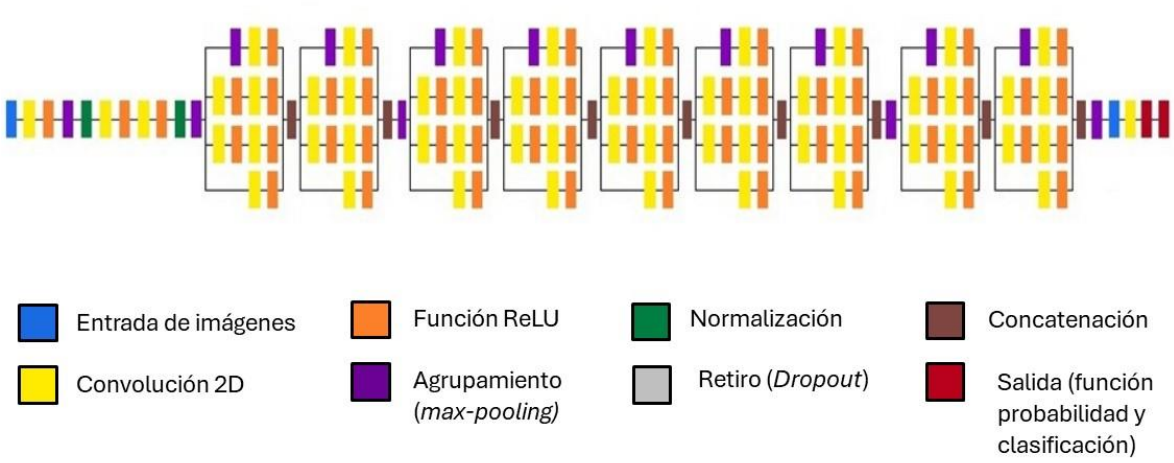

Fig. 10. Estructura de red GoogleNet.

red cuenta con una secuencia de nueve *clusters*, los cuales están compuestos de cuatro subestructuras con capas convolutivas, de agrupamiento o *pooling* y funciones ReLU organizadas en paralelo, cuya salida (bloques color café) concatena las cuatro subestructuras

para pasar al siguiente *cluster* o función. En la etapa final o de salida, la red cuenta con una capa de retiro o *dropout* (bloque gris), una función de activación sigmoide y una función de pérdida *crossentropy* (ambas en color rojo).

Las capas finales de la red se modificaron para adaptarlas al entrenamiento de los nuevos datos. A partir de la capa 140 se agregaron las capas de la **Tabla IV**, para obtener la estructura de red deseada. La salida de la red comprende una función sigmoide, realizando la clasificación en dos categorías, i.e., SR normal o SR con sibilancias. Así mismo, se realizó

una regularización del modelo para evitar el sobreajuste de la red al añadir una capa *dropout* de 40%. Además, se trabajó con varias modificaciones en las últimas capas de la red para depurarla y ajustar los hiperparámetros como se menciona a continuación.

Se realizó una búsqueda exhaustiva de hiperparámetros, para seleccionar aquellos que ofrecieran mejores resultados de exactitud en la clasificación para cada TFR, empleando como conjunto de búsqueda los parámetros mostrados en la **Tabla V**, con distintas opciones para los parámetros número de épocas, tamaño de lote, regularización L2, frecuencia de

Tabla IV. Capas agregadas a red pre-entrenada GoogleNet

Número de capa	Tipo	Tamaño
140	*Average pooling*	1 x 1 x 1024
141	*Dropout* (40%)	1 x 1 x 1024
142	*Fully connected*	1 x 1 x 2
143	Sigmoide	1 x 1 x 2
144	Clasificación (*crossentropy*)	1 x 1 x 2

validación y tasa de aprendizaje inicial. La elección de estos valores está basado en los parámetros empleados en trabajos que emplean la misma base de datos ICBHI Challenge,

e.g.[31], [32], [33], [23], además de parámetros empleados experimentalmente en este trabajo en las pruebas preliminares mediante los cuales se obtuvieron buenos resultados. Cabe mencionar que, de esta combinación se obtienen 405 combinaciones de hiperparámetros, de los cuales se seleccionó el que ofrece los mejores resultados de exactitud en entrenamiento y validación, para cada una de las cuatro TFR. Este método fue seleccionado para aprovechar el conocimiento previo de que la combinación de parámetros de la **Tabla V** obtuvo los resultados más sobresalientes al realizar las pruebas preliminares. Sin embargo, otras técnicas como optimización bayesiana o búsqueda aleatoria se pueden plantear para un trabajo futuro.

Sobre los conjuntos de entrenamiento, validación y prueba, se utilizó 70% de los datos para el conjunto de entrenamiento y 25% para el conjunto de validación, además de contar con un bloque de prueba nunca vista por la red, el cual equivale al 5% de los datos. Cabe mencionar, que los sujetos de cada conjunto (entrenamiento, validación y prueba) son distintos entre sí. Además, considerando la cantidad de datos con la que se cuenta, se realizó un procedimiento de validación cruzada con *k*-dobleces (k=5) para realizar la evaluación del desempeño del clasificador, donde para cada partición *i*, se entrenó el modelo con las con *k*-1 particiones restantes, para después evaluarlo en el conjunto i. La calificación del modelo se obtiene al calcular el promedio de las *k* evaluaciones individuales. Cabe mencionar que la mayoría de los trabajos encontrados en la revisión del estado del arte no reportan resultados sobre un conjunto de prueba, ya que solo emplean conjunto de entrenamiento y validación,

o en ocasiones emplean parte de los datos del conjunto de entrenamiento como conjunto de prueba, lo que pudiera conducir a resultados de comprobación no concluyentes.

En la **Fig. 11** se muestra un diagrama de bloques que contiene los pasos que a seguir para realizar la detección de ciclos respiratorios normales y ciclos que contienen sibilancias, donde en el panel izquierdo se observa la etapa de preprocesamiento de señales, a partir de las cuales se obtienen las cuatro distintas TFR, para posteriormente fungir como entrada a la red neuronal pre-entrenada *GoogleNet* y realizar la detección.

IV.5.2. Índices de desempeño

El desempeño del detector automático de sibilancias se basó en la información de las anotaciones de eventos en la base de datos ICBHI realizadas por los expertos. La anotación de eventos es el método más común para evaluar la robustez de algoritmos para detectar SR

Tabla V. Valores de búsqueda exhaustiva de hiperparámetros

Número de épocas	Tamaño de lote	Regularización L2	Frecuencia de validación	Tasa de aprendizaje inicial
50	64	1×10^{-1}	5	1×10^{-2}
100	128	1×10^{-5}	10	1×10^{-3}
150	256	1×10^{-10}	30	5×10^{-3}
				3×10^{-4}
				1×10^{-5}

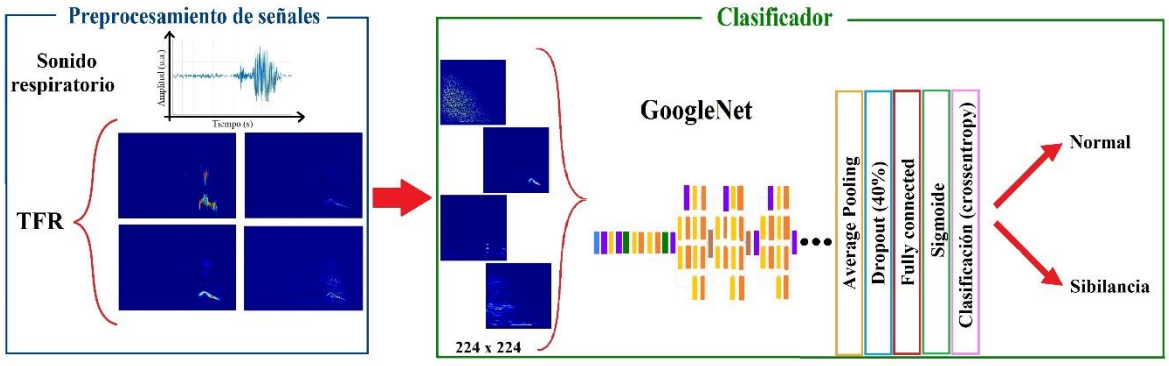

Fig. 11. Descripción general del método propuesto.

adventicios [21]. De manera cuantitativa, se evaluó el desempeño del clasificador de acuerdo con los parámetros de exactitud (**Ec. 29**),

$$Exactitud = \frac{TP+TN}{TP+FP+TN+FN} \quad , \tag{29}$$

sensibilidad (**Ec. 30**)

$$Sensibilidad = \frac{TP}{TP+FN} \ , \tag{30}$$

especificidad (**Ec. 31**)

$$Especificidad = \frac{TN}{TN+FP} \ , \tag{31}$$

precisión (**Ec. 32**)

$$Precisión = \frac{TP}{TP+FP} \quad , \tag{32}$$

y F_1-Score (**Ec. 33**)

$$F_1 - Score = \ 2\ x \frac{\text{Precisión}\, x\ Sensibilidad}{Precisión+Sensibilidad} \tag{33}$$

de la clasificación, donde TP es el resultado de prueba que indica correctamente los ciclos respiratorios normales, TN indica los ciclos con sibilancias que fueron detectados

satisfactoriamente, FP es el resultado de prueba que indica la presencia errónea de ciclos respiratorios normales y FN indica los ciclos respiratorios con sibilancias identificados incorrectamente. Además, se incluye el parámetro F_1-*Score* como una medida que incorpora un valor medio entre precisión y sensibilidad. Estos índices se evalúan en un rango de 0 a 1, donde un valor más alto es considerado mejor o con mayor eficacia de desempeño del detector.

Capítulo 5

Resultados y discusión

En esta sección se presentan los resultados obtenidos al aplicar los distintos métodos de TFR, i.e., SP, SST, HHS y WVD, a la señal sintética simulada, así como a los ciclos respiratorios reales de la base de datos ICBHI Challenge 2017. Las TFR se emplean de forma separada como entrada a la red neuronal *GoogleNet* modificada para realizar su entrenamiento de clasificación en SR normales o SR con sibilancias. Además, se presentan los resultados alcanzados con los hiperparámetros seleccionados, para posteriormente compararlos con los resultados de clasificación encontrados en la literatura para esta misma base de datos.

V.1. Sibilancia simulada y su TFR ideal

La señal obtenida a partir de la **Ec. 27** se presenta en la **Fig. 12**, donde se observan los segmentos de señal monofónica al inicio y final de la señal, entre 0-0.025 s y 0.275-0.3 s, así como el componente polifónico en la parte central, entre 0.025 y 0.275 s. Esta señal representa un sonido adventicio continuo polifónico que cumple con las características de una sibilancia. A partir de la señal sintética simulada, mostrada en la **Fig. 12**, se calcularon sus componentes de frecuencia instantánea, mediante la derivada de la fase de la expresión de la sinusoidal modulada en frecuencia, así como el componente de amplitud instantánea, empleando el módulo de la relación de Euler con el componente real y el imaginario de la señal. De esta manera, se obtuvo la TFR ideal teórica mostrada en la **Fig. 13**, la cual permite visualizar los distintos componentes presentes en la señal de sonido adventicio continuo polifónico, donde desde el tiempo 0 a 0.3 s se observa el componente monofónico de frecuencia modulado sinusoidalmente que oscila alrededor de 75 a 90 Hz, además del

componente monofónico de frecuencia modulado linealmente entre 0.025 y 0.275 s que aparece entre 160 y 230 Hz. Esta TFR ideal teórica permite tener un patrón para la posterior comparación de ésta con las TFR obtenidas al aplicar los distintos parámetros pertinentes a cada una.

V.1.1. Distribución Wigner-Ville

La WVD, como fue mencionado en el apartado III.3.2 referente a sus características, es una distribución única para cada señal, por lo que no requiere de la selección de parámetros para

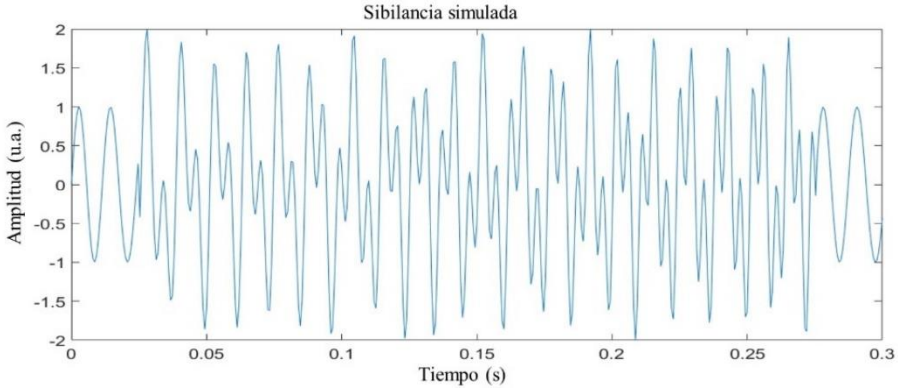

Fig. 12. Señal de sonido adventicio continuo polifónico sintético.

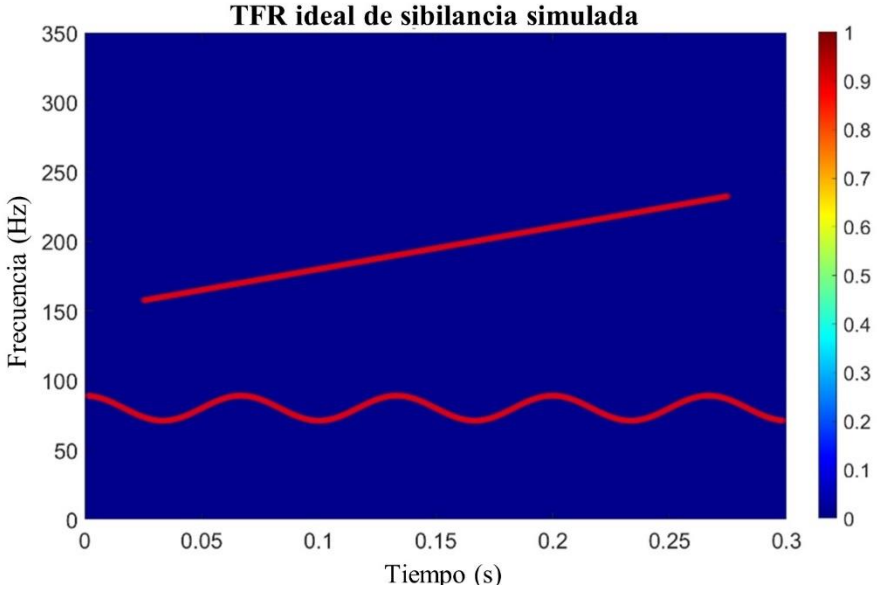

Fig. 13. TFR ideal de la sibilancia simulada.

ser realizada. En la **Fig. 14** se muestra la WVD correspondiente a la señal de sonido adventicio continuo polifónico. Se puede observar el componente de frecuencia modulado linealmente, sin embargo, en los extremos hacia 0.025 y 0.275 s alrededor de 170 y 230 Hz respectivamente, comienza a aparecer disperso o difuminado.

De la misma manera, el componente de frecuencia modulado sinusoidalmente de la parte inferior de la figura aparece con baja resolución en la frecuencia, además de presentar líneas difusas hacia los extremos en 0 y 0.3 s. Una de las características más importantes de este tipo de TFR, es la aparición de los términos cruzados, los cuales aparecen como un conjunto de líneas curveadas ubicadas a medio camino entre los términos de frecuencia originalmente presentes en la señal, mencionados anteriormente.

V.1.2. Espectrograma

De acuerdo con los parámetros del SP, se realizaron distintas pruebas al aplicar diferentes tipos y longitudes de ventana, manteniendo el traslape entre segmentos del 50%. En la **Fig. 15** se muestran los resultados de este procedimiento, donde en el panel superior izquierdo

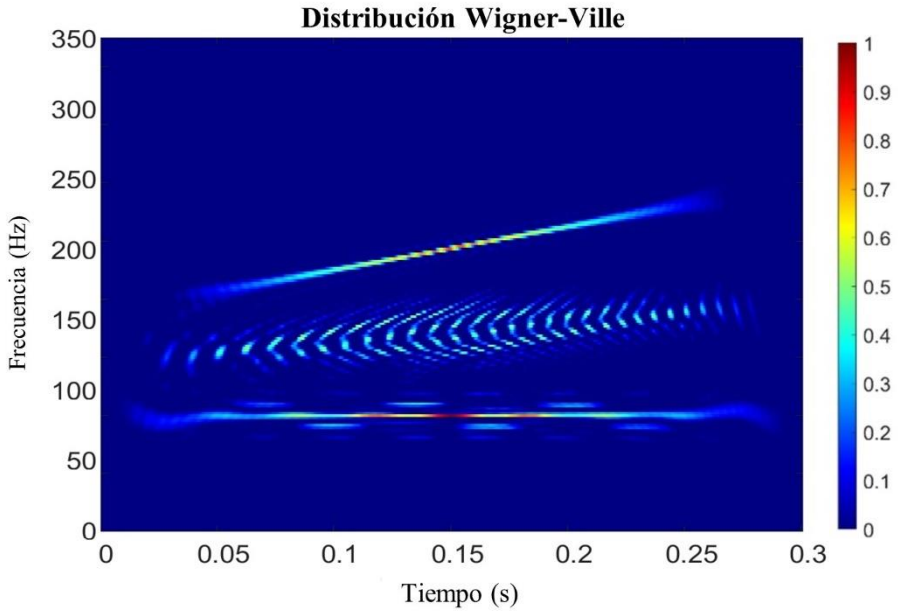

Fig. 14. WVD de la sibilancia simulada.

(**Fig. 15A**) se muestra el SP obtenido al emplear una ventana Hanning de 15 ms, la cual al emplear una ventana tan corta resulta en una buena resolución temporal, pero una pobre resolución espectral, ya que se aprecian segmentos de alta energía alargados en la frecuencia, que aparecen como líneas verticales entre ambos componentes principales de la señal.

En el panel superior derecho (**Fig. 15B**) se muestra el SP realizado con una ventana Hamming de 32 ms, la cual aparece con una mejor resolución en la frecuencia, si lo comparamos con la **Fig. 15A**, ya que los componentes de energía más alta, mostrados en color rojo, corresponden a los elementos principales de la señal, los modulados en frecuencia lineal y sinusoidalmente. A pesar de que la forma de ambos componentes es notablemente más nítida, este SP aún presenta una baja resolución en la frecuencia. Además, aún son notables los términos cruzados presentes en el SP, como en el caso de la **Fig. 15A**. La **Fig. 15C** muestra el SP realizado con una ventana Blackman de 100 ms, cuya longitud de ventana

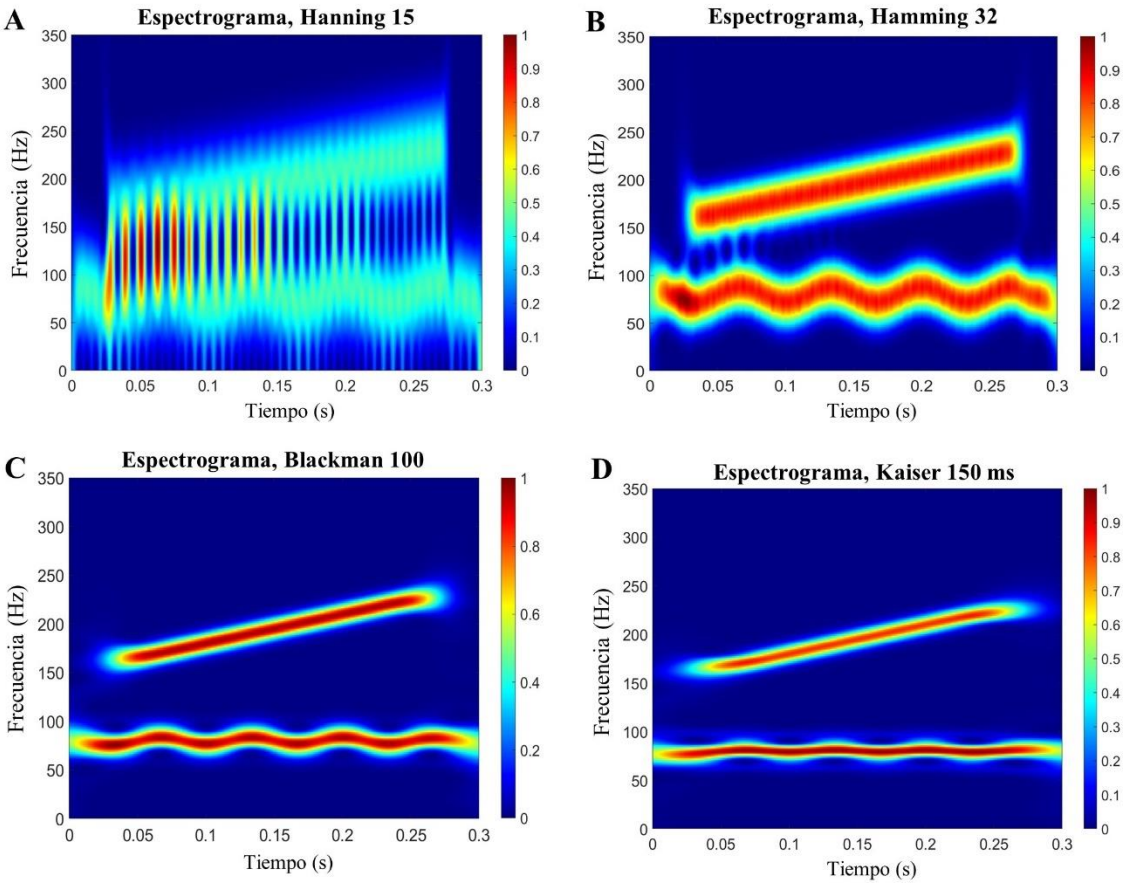

Fig. 15. SP con distintos parámetros de la sibilancia simulada.

es comúnmente empleada para el procesamiento y detección de sibilancias, debido a sus características [9]. Además, esta TFR permite la visualización más definida de ambos componentes, tanto en el tiempo como en la frecuencia. Mientras que, en el último panel, **Fig. 15D**, al emplear una ventana Kaiser de 150 ms, se observa una mejor resolución en la frecuencia, pero una resolución más baja en la escala temporal, ya que sus componentes se observan muy alargados horizontalmente.

V.1.3. Espectro de Hilbert-Huang

A la señal obtenida de la **Ec. 28** se le aplicó el método de EMD clásico para obtener los IMF de la señal. Como resultado se muestra la **Fig. 16**, donde se presentan 3 de los 6 IMF obtenidos a partir de esta sibilancia polifónica simulada. En el primer renglón se muestra la señal original de la **Fig. 12**, mientras que en el segundo renglón se muestra el primer IMF obtenido a partir de la descomposición, cuyos componentes corresponden a los de mayor

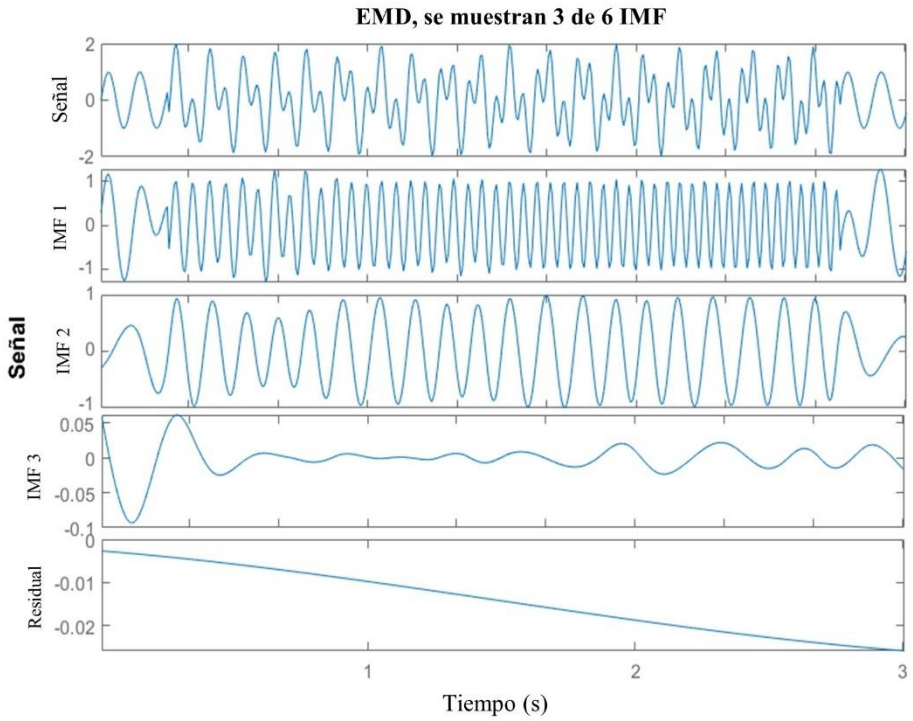

Fig. 16. Descomposición EMD clásica de sibilancia simulada.

frecuencia y amplitud de la señal. En el tercer y cuarto renglón se muestran el segundo y tercer IMF, respectivamente, donde se visualiza la reducción de amplitud y frecuencia conforme avanza el procedimiento. En el último apartado aparece la señal residual, de la cual ya no se puede obtener otro IMF al no cumplir con la característica mencionada anteriormente de los cruces por cero y los mínimos y máximos locales.

Como muestra de que estos IMF cumplen con las características propias de estas funciones, en la primera columna de la **Tabla VI** se menciona el número de cruces por cero que contiene cada IMF, en la segunda columna el número de extremos y, en la última columna la diferencia de cruces por cero y extremos, la cual no debe exceder de uno para que un componente sea considerado IMF. En la **Fig. 17** se presenta la comparación de dos métodos de descomposición, donde se puede apreciar cómo es que cada método, EMD izquierda y CEEMDAN derecha, realizan la descomposición de maneras diferentes a pesar de tener como entrada la misma señal simulada como sibilancia polifónica. Se aprecian diferencias desde el IMF 1, en el método CEEMDAN el IMF aparece con una amplitud más constante, en comparación con el método EMD para el mismo modo, el cual aparece con picos de mayor amplitud.

Lo mismo sucede con el modo 2 y 3, donde al eliminar ciertos componentes que están incluidos en los primeros IMF las señales disminuyen gradualmente su amplitud y frecuencia. Otra diferencia se puede observar en la señal residual, donde en el método EMD este residuo es casi un componente recto, mientras que en CEEMDAN presenta una curvatura y su amplitud es menor que en EMD.

Tabla VI. Características de los IMFs obtenidos mediante EMD para la sibilancia simulada

IMF	Número de cruces por cero	Número de extremos	Número de cruces por cero – número de extremos
1	107	108	-1
2	45	45	0
3	21	20	1
4	8	8	0
5	4	3	1
6	2	1	1

A partir del análisis anterior se obtuvieron las imágenes TF de la **Fig. 18**, las cuales muestran dos distintos métodos de descomposición, donde la **Fig. 18A** muestra el método clásico EMD, mientras que la **Fig. 18B** muestra el HHS obtenido al realizar descomposición CEEMDAN. Al comparar ambos métodos con la TFR ideal de la **Fig. 12** se puede observar que la EMD clásica genera una TFR cuyos componentes de frecuencia modulada linealmente y modulada sinusoidalmente aparecen con potencia más baja que los generados por CEEMDAN. Además, el componente de frecuencia modulada linealmente en la **Fig. 18B** se muestra con menos oscilaciones que el realizado con EMD, siendo más cercano a la TFR

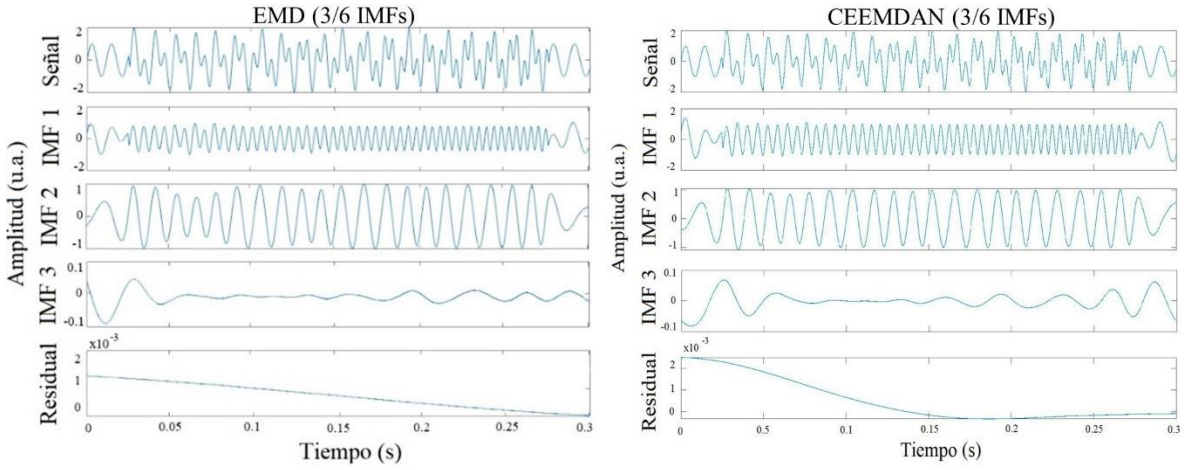

Fig. 17. Descomposición en IMFs de la sibilancia simulada. Izquierda: EMD, derecha: CEEMDAN.

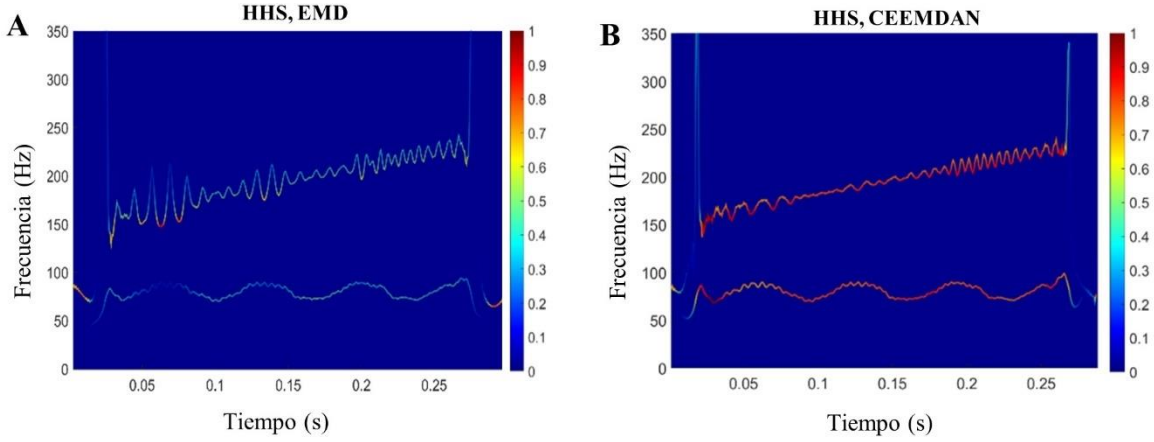

Fig. 18. Comparación de HHS mediante EMD y CEEMDAN.

ideal. En la **Fig.19** se muestran los HHS obtenidos a partir de variar el número de IMF utilizados en la TFR, empleando el método CEEMDAN, donde la **Fig. 19A** muestra la TFR obtenida al utilizar únicamente el primer IMF, el cual corresponde al componente de frecuencia modulada linealmente, con variaciones de energía y marcadas oscilaciones de frecuencia. Posteriormente, en la **Fig. 19B** se muestra el HHS obtenido al emplear únicamente el segundo IMF obtenido a partir de CEEMDAN, esta TFR muestra el componente de frecuencia modulada sinusoidalmente, donde las frecuencias más altas de este componente aparecen con energía menor que las frecuencias más bajas. Finalmente, en la **Fig. 19C** aparece el HHS obtenido al utilizar los dos IMF producidos en la descomposición CEEMDAN, el cual muestra que ambos componentes de frecuencia cuentan con mayor energía, además de ser más estable, i.e., con energía mayor a 0.8. Además, se observan menos oscilaciones en los componentes, si lo comparamos con las imágenes realizadas con IMF separados.

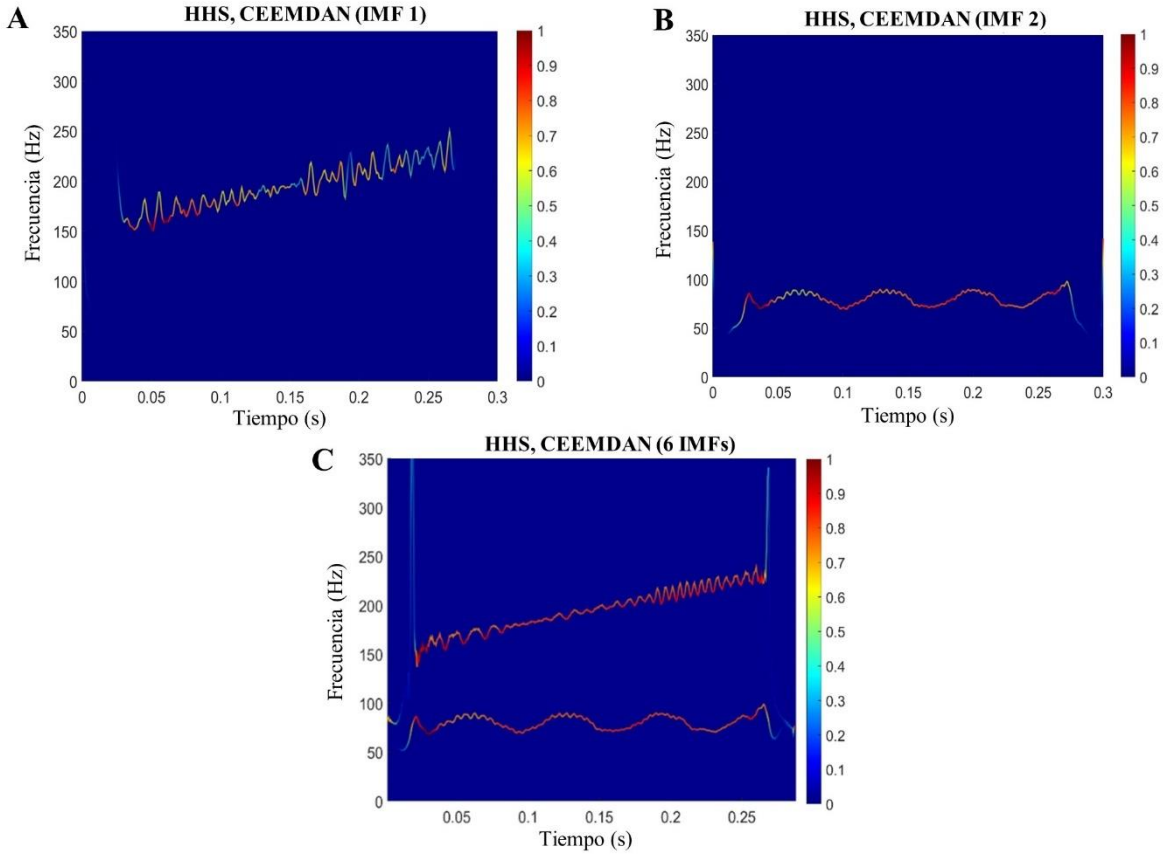

Fig. 19. Comparación de HHS – CEEMDAN, al emplear diferentes IMFs.

V.1.4. Transformada Synchrosqueezing

Para la técnica TFR de SST se realizó la comparación entre distintos parámetros: 1) tipo de ventana, y 2) longitud de ventana. Los resultados obtenidos se presentan en la Fig. 20, donde la **Fig. 20A** muestra la SST realizada al emplear una ventana Hamming de 32 ms, donde se observa una resolución muy pobre tanto en tiempo como en frecuencia, donde ambos componentes aparecen con grandes saltos en intervalos de frecuencia. En la **Fig. 20B** se muestra la SST obtenida al emplear una ventana Kaiser de 142 ms, la cual aparece con mejor resolución que la primera, sin embargo, aún resulta muy pobre y alejada de la TFR ideal, ya que muestra componentes difusos en tiempo y frecuencia para ambos componentes, además de mostrar potencias más bajas que la representación ideal de la misma señal. En la **Fig. 20C** aparece la representación realizada a partir de una ventana Hanning de 142 ms, en esta imagen podemos apreciar la diferencia que existe al seleccionar otro tipo de ventana manteniendo la longitud de la figura anteriormente descrita. En esta TFR la ventana Hanning crea una representación con componentes más estrechos y definidos en ambas dimensiones TF que la realizada con la ventana Kaiser, sin embargo, aún se encuentra alejada de la representación ideal de la **Fig. 12**. Finalmente, en la **Fig. 20D**, se muestra la TFR realizada con una ventana Blackman de 100 ms, longitud que ha sido muy utilizada en la literatura para realizar análisis de sibilancias, debido a las características de estas. La representación obtenida con estos parámetros muestra los componentes con mayor definición de esta comparación, ya que es fácil apreciar visualmente las características de ambos componentes de frecuencia modulados lineal y sinusoidalmente, además, la energía es más constante y con menores variaciones como en las representaciones anteriores.

V.2. Selección de parámetros de TFR

Se empleó el método de información mutua (*mutual information*, MI), para realizar una comparación cuantitativa entre las distintas TFR y sus parámetros, de manera que no solo sea visual la elección de las ventanas para el SP y SST, y del método de descomposición para el HHS. El método MI calcula una cantidad o medida de dependencia entre dos variables

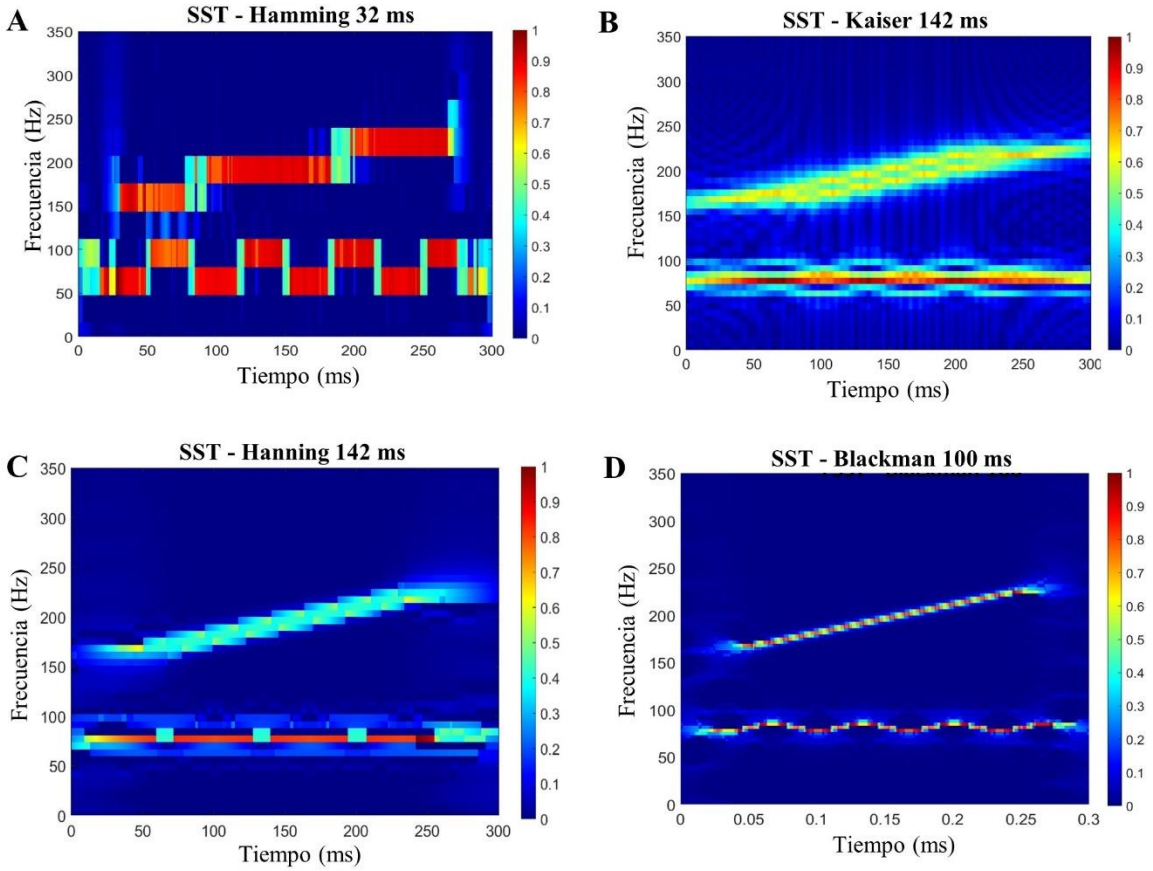

Fig. 20. Comparación de SST al emplear distintos parámetros.

aleatorias, el valor calculado siempre es igual o mayor a cero y mientras más alto sea el valor mayor es la relación entre ambas variables [54]. Este método es ampliamente utilizado en análisis de imágenes, por lo que se empleó para cuantificar la similitud entre distintas TFR. El análisis MI se realizó mediante la función del *software* MATLAB, implementada por Delpiano [55], la cual emplea una función cuyas entradas son dos imágenes de tamaños iguales.

Para realizar el cálculo de MI se utilizó una señal simulada con un componente de frecuencia modulado sinusoidalmente que corresponde al factor obtenido mediante la **Ec. 28**. En la **Fig. 21** se muestran las distintas representaciones de SP obtenidas empleando el tipo de ventana Blackman al utilizar diversas longitudes desde 15 hasta 200 ms, así como su

índice de similitud mediante el análisis MI, que aparece debajo de cada una de ellas, estos valores también se muestran en la **Tabla VII**.

De la misma manera, en la **Fig. 22**, se muestra el análisis mediante MI, para tres representaciones de HHS, realizadas mediante CEEMDAN, ya que este método de descomposición es el que brinda mayores beneficios para realizar la TFR de HHS. La primera es la representación que emplea un solo IMF, la segunda emplea dos de los tres IMF obtenidos a partir de CEEMDAN, mientras que el último panel emplea los tres IMFs calculados con este método. Conforme incrementa el número de funciones empleadas, la TFR aparece más completa, al aumentar la información contenida en los IMF. En la parte

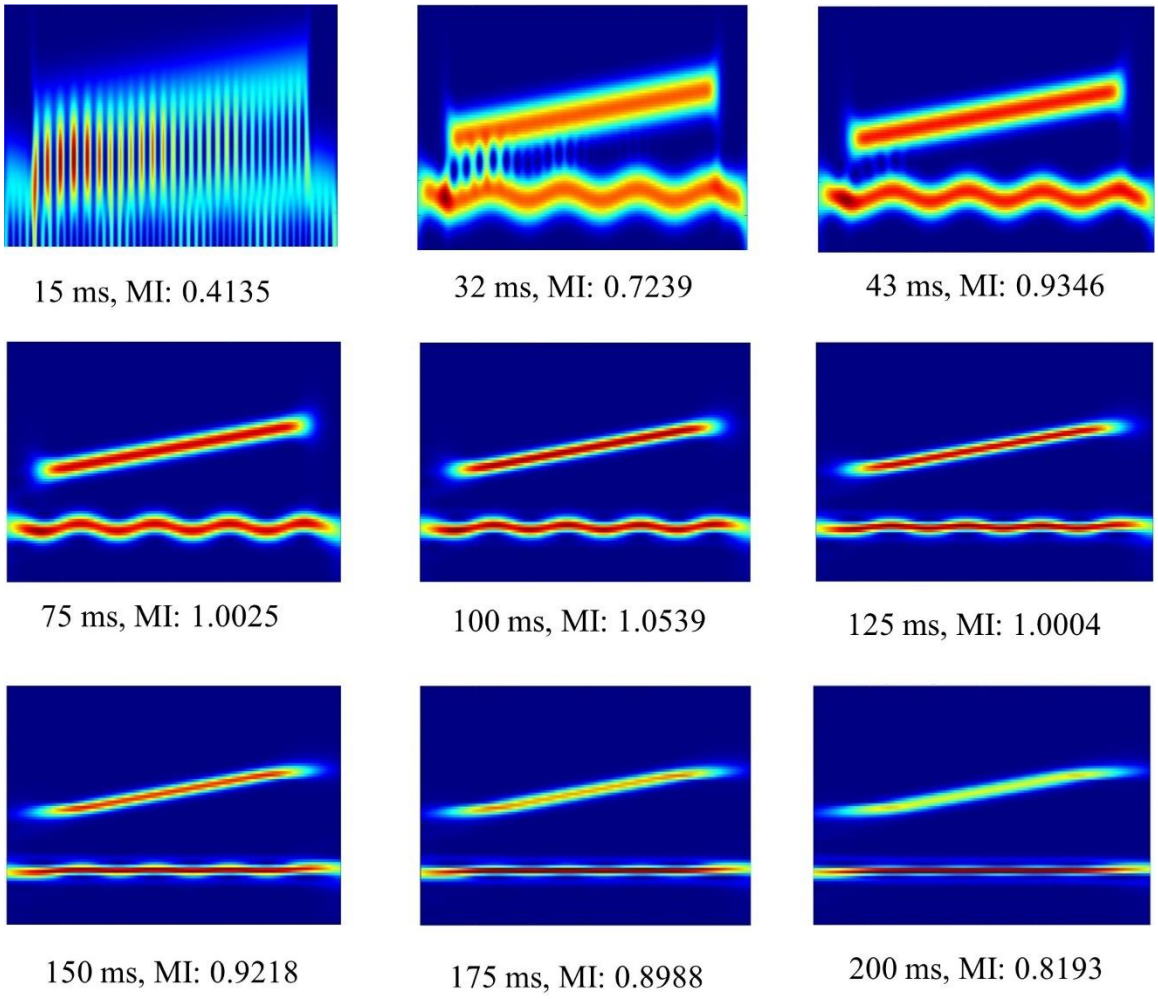

Fig. 21. Comparación de índice MI en SP, empleando distintos parámetros.

inferior de las imágenes se muestra el número de IMF empleados al calcular el espectro y su índice MI.

El mismo método se realizó para obtener la similitud entre las representaciones obtenidas mediante diferentes longitudes de ventana para la SST, empleando el tipo Blackman. A manera de ejemplo, en la **Fig. 23** se muestran las distintas TFR obtenidas mediante esta ventana, al variar la longitud de esta entre 15 y 200 ms, así como su índice MI, cuyos valores obtenidos también se muestran en la **Tabla VII**.

Finalmente, se seleccionaron como parámetros para las distintas TFR: 1) ventana Blackman de 100 ms para el SP, 2) emplear todos los IMF obtenidos a partir de la descomposición CEEMDAN para el HHS, y 3) ventana Blackman de 100 ms para la SST. La WVD no requirió de análisis debido a su característica de no contar con parámetros modificables que puedan repercutir en la distribución TF, lo cual puede considerarse una gran ventaja, ya que este método no depende de la elección de los parámetros como para las otras TFR.

Tabla VII. Valores obtenidos al evaluar el índice de información mutua (MI) en distintas longitudes de ventana tipo Blackman para SP y SST.

	Longitud de ventana (ms)								
TFR	**15**	**32**	**43**	**75**	**100**	**125**	**150**	**175**	**200**
SP	0.4135	0.7239	0.9346	1.0025	1.0539	1.0004	0.9218	0.8988	0.8193
SST	0.6920	0.8993	0.9895	1.1396	1.1443	0.9018	0.8674	0.8323	0.7980

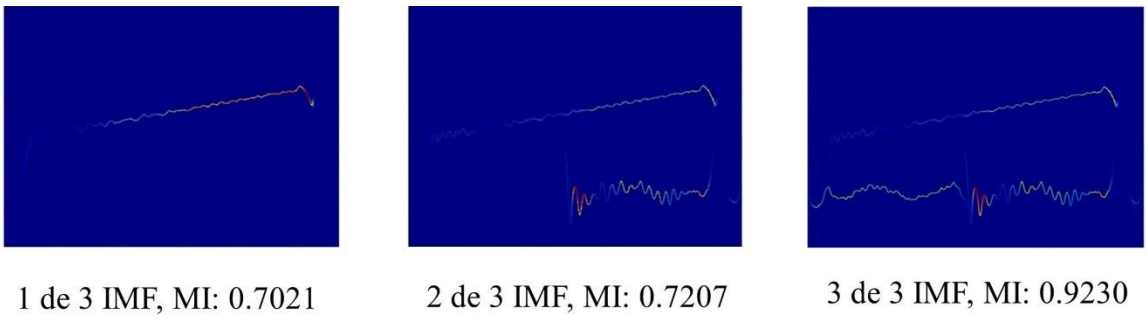

Fig. 22. Comparación de índice MI para HHS empleando distintos parámetros.

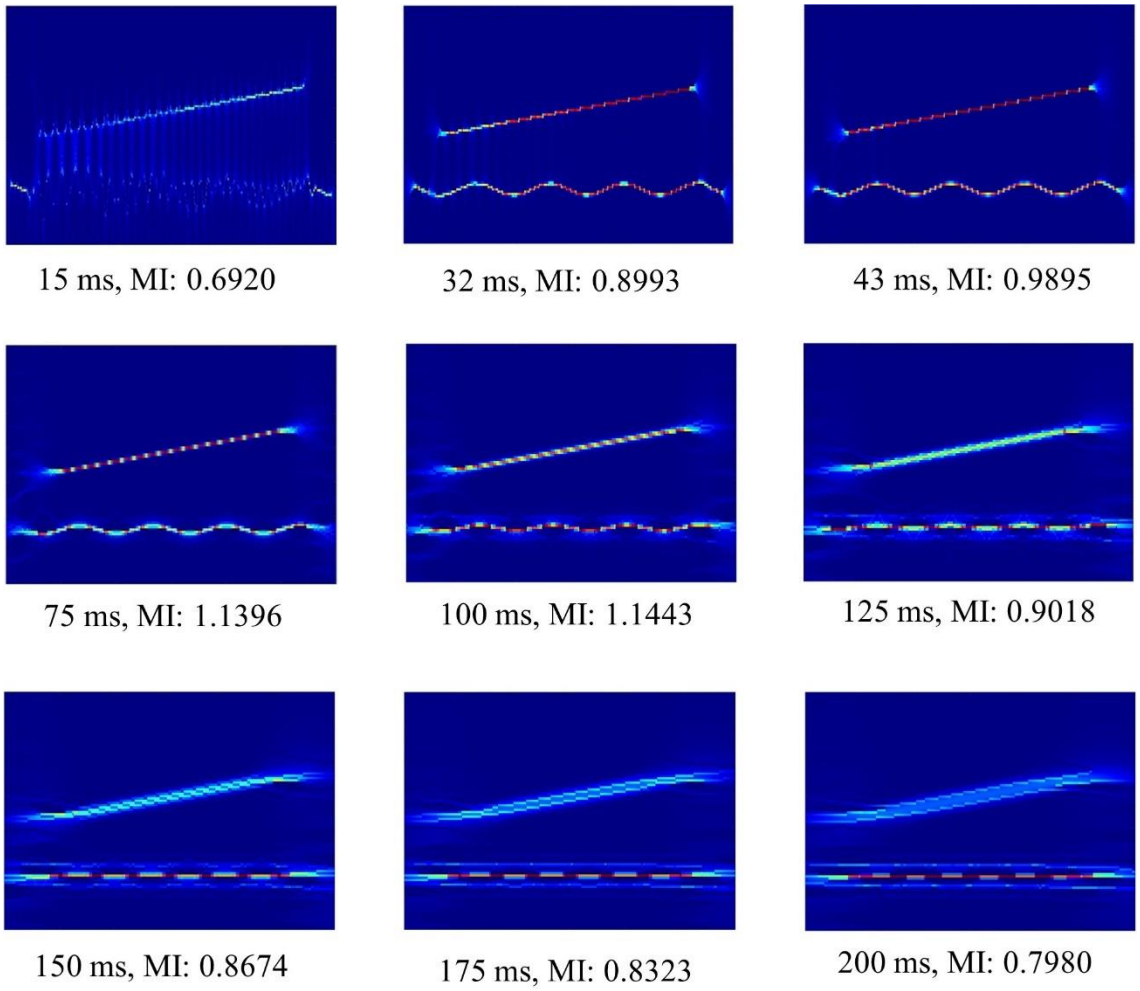

Fig. 23. Comparación de índice MI para SST, empleando distintos parámetros.

En la **Fig. 24A** se muestra la TFR ideal de la sibilancia simulada, para visualizarla como punto de comparación para las otras representaciones. Se calculó la WVD de la sibilancia simulada, mostrada en la **Fig. 24B**. Posteriormente, se aplicó a la técnica TF clásica, SP, una ventana Blackman de 100 ms, mostrada en la **Fig. 24C**. Así como la SST basada en Fourier, utilizando una ventana de las mismas características la cual se muestra en la **Fig. 24D**. Por último, en la **Fig. 24E** se muestra la señal con HHS aplicando el método CEEMDAN para la descomposición de modos, empleando todos los IMF resultantes. Para desplegar las TFR se utilizó el mismo mapa de colores, la misma escala de frecuencia (Hz) en el eje *y*, y la misma escala de tiempo (s) en el eje *x* para facilitar la comparación entre técnicas. Cabe mencionar que se seleccionaron los parámetros que ofrecen la mejor

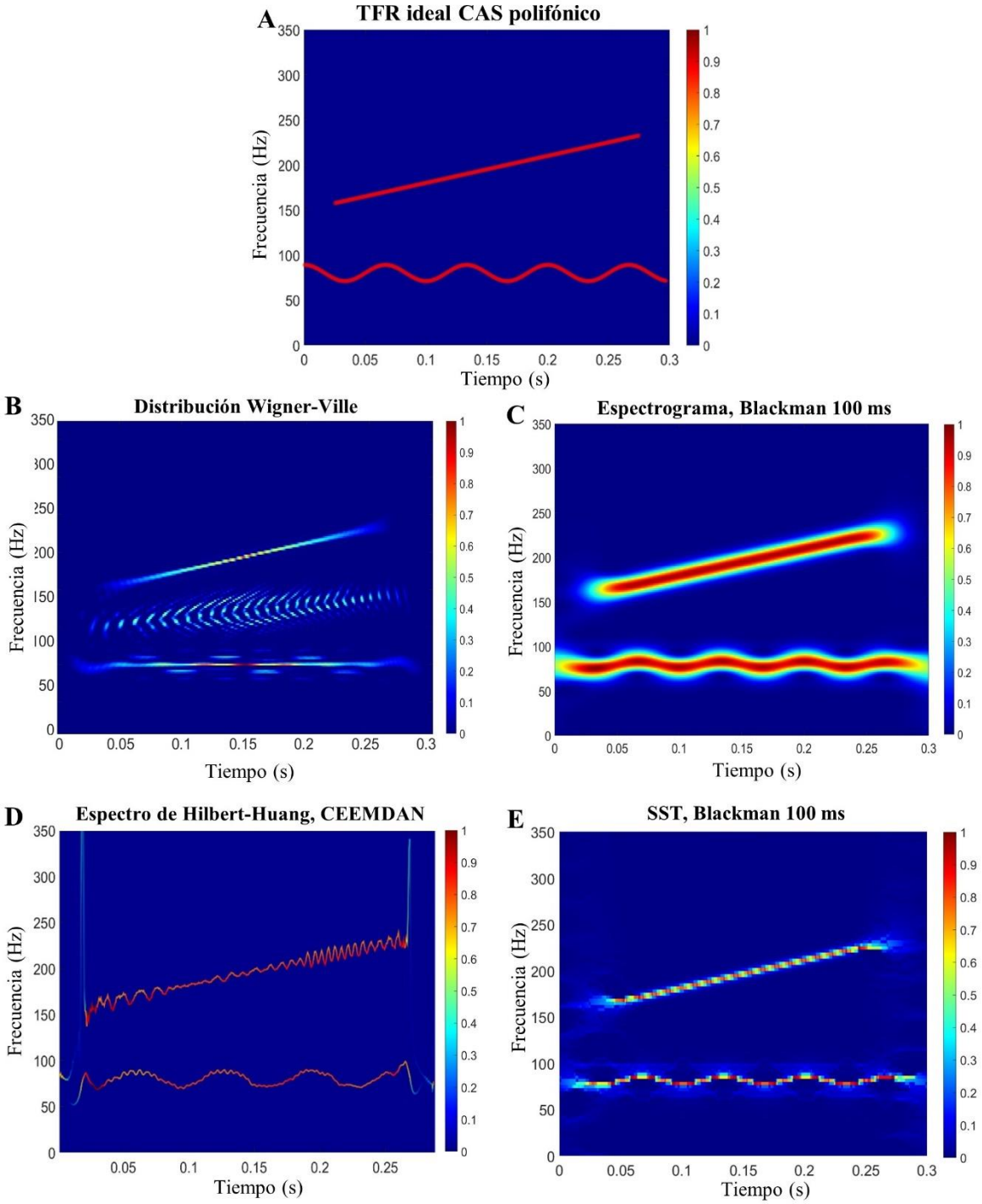

Fig. 24. TFRs seleccionadas para la sibilancia simulada. A) TFR ideal de sonido adventicio polifónico, B) WVD, C) SP con ventana Blackman de 100 ms, D) HHS empleando el método CEEMDAN y D) SST con ventana Blackman de 100 ms

representación de cada categoría. En esta figura se pueden apreciar las diferencias entre las distintas TFR, su similitud con la representación ideal (**Fig. 24A**) y la dispersión del eje

temporal y el eje espectral, lo que permite visualizar las ventajas y desventajas de cada método. Puede observarse que la SST es la TFR que ofrece mejor resolución tanto temporal como espectral para la sibilancia polifónica sintética.

V.3. Sonidos respiratorios reales de base de datos ICBHI

Se aplicó el preprocesamiento descrito anteriormente y las cuatro distintas técnicas TF con parámetros seleccionados a señales obtenidas de la base de datos ICBHI, como la señal mostrada en la **Fig. 25** perteneciente a un ciclo respiratorio normal.

V.3.1. Ejemplo de ciclo respiratorio con SR normal

En la **Fig. 26A** se muestra la WVD de un SR normal perteneciente a la señal de la **Fig. 25**, donde se observa una dispersión de energía que pareciera una nube de puntos, cuya concentración pudiera aumentar debido a los términos cruzados de la propia representación. Mientras que la **Fig. 26B** presenta el SP obtenido al emplear una ventana Blackman de 100 ms a la misma señal de SR normal, donde se observa un componente de mayor energía cerca de 0.41 s, el cual puede deberse a la presencia de un sonido cardiaco. En el tercer panel, en la **Fig. 26C**, se muestra el HHS del mismo segmento utilizando el método de descomposición CEEMDAN, empleando todos los IMF calculados. En esta TFR resalta el componente ubicado cerca de 0.14 s, el cual parece corresponder al componente de mayor energía en el mismo instante de tiempo, encontrado alrededor de 600Hz. Al comparar ambas representaciones podemos observar sus diferencias, ya que el HHS presenta componentes verticales, alargados en la frecuencia, mientras que el SP se dispersa mayormente a lo largo del tiempo. Finalmente, en la **Fig. 26D** se muestra la SST del mismo ciclo respiratorio obtenida con el mismo tipo y longitud de ventana que para el SP, Blackman de 100 ms. En esta figura se puede apreciar que la SST (**Fig. 26D**) presenta los mismos componentes que aparecen en el SP (**Fig. 26B**) pero cuentan con una mayor resolución en ambos ejes temporal y espectral. Además, en esta misma TFR se pueden apreciar líneas horizontales delgadas de longitud y energía variables, que en el SP se muestran más tenues. Esta técnica de alta resolución pudiera permitir una mejor identificación de los SR y su clasificación, incluso a

simple vista. En la SST no se aprecian a simple vista componentes grandes con alta energía (en color rojo) ya que, debido a su resolución, los componentes de alta energía del SP pudieran aparecer como puntos muy delgados en la SST. Para las TFR se utilizó el mismo mapa de colores, la misma escala de frecuencia (Hz) en el eje-*y* y el mismo segmento de tiempo (s) en el eje-*x* para facilitar la comparación entre técnicas.

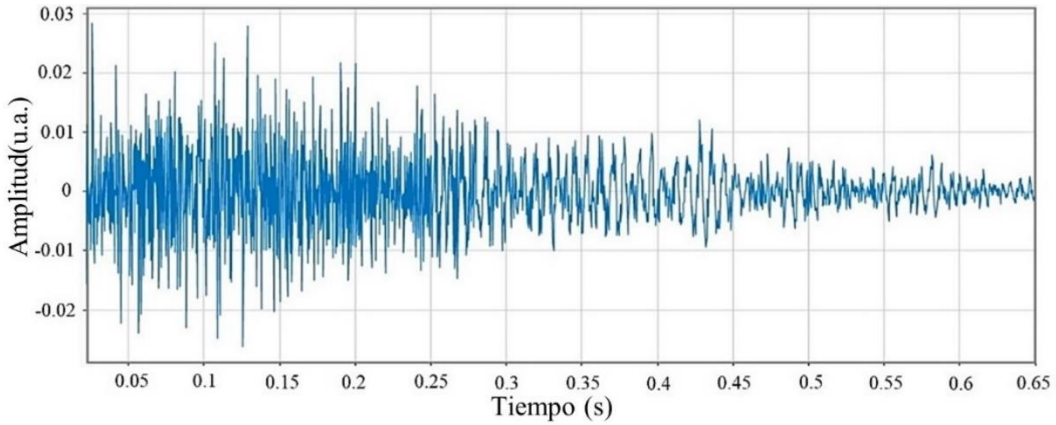

Fig. 25. Señal de un ciclo de SR normal.

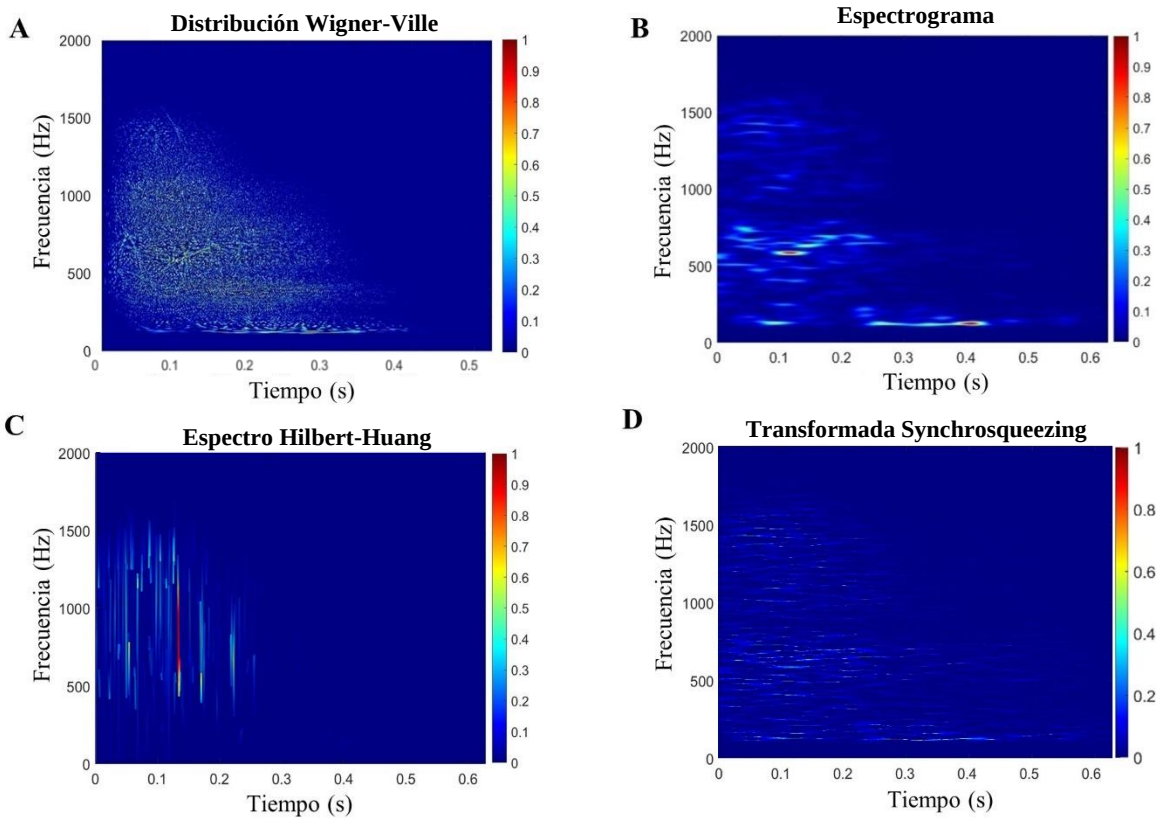

Fig. 26. TFRs de ciclo de SR normal.

V.3.2. Ejemplo 1 de ciclo respiratorio con SR con sibilancia

En otro ejemplo de la base de datos ICBHI, se muestra la señal de un ciclo respiratorio con sibilancias en la **Fig. 27**. A partir de la cual se obtuvieron las TFR mostradas en la **Fig. 28**, donde para cada TFR obtenida se resalta el rango de frecuencia de la sibilancia con líneas rojas punteadas. La **Fig. 28A** muestra la WVD, donde aparecen líneas cuasi-horizontales entre 0.8 y 1.1 s, entre 100 y 290 Hz, que indican la presencia de una sibilancia. En la **Fig. 28B** se muestra el SP obtenido al utilizar una ventana Blackman de 100 ms, donde se observa la misma sibilancia, pero más difusa en la dimensión espectral y temporal. En la **Fig. 28C** se muestra el HHS empleando el método CEEMDAN, donde se distinguen los componentes en frecuencia de la sibilancia observada en la **Fig. 28A**. Sin embargo, el componente de menor frecuencia, alrededor de 200 Hz pareciera recortado o atenuado en la parte central, debido a que la mayor energía en el tiempo 0.65 s se encuentra cerca de los 550 Hz. Finalmente, en la **Fig. 28D,** se presenta la SST del ciclo respiratorio con sibilancia, al emplear la misma ventana del SP. La SST mantiene una gran similitud con el SP mejorando la resolución en tiempo y frecuencia, debido a la delgadez de los componentes se aprecia más tenue, sin embargo, al realizar un acercamiento se aprecian los detalles de alta resolución de esta TFR.

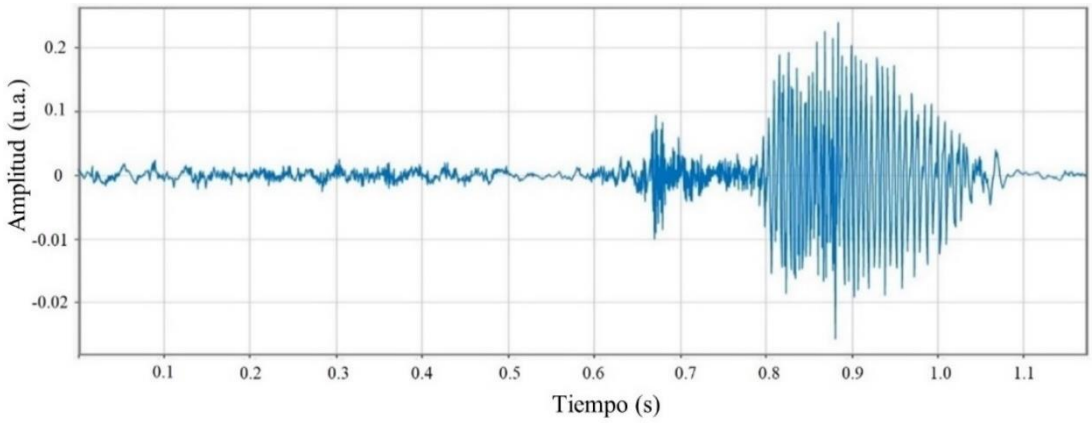

Fig. 27. Señal de un ciclo de SR con sibilancia.

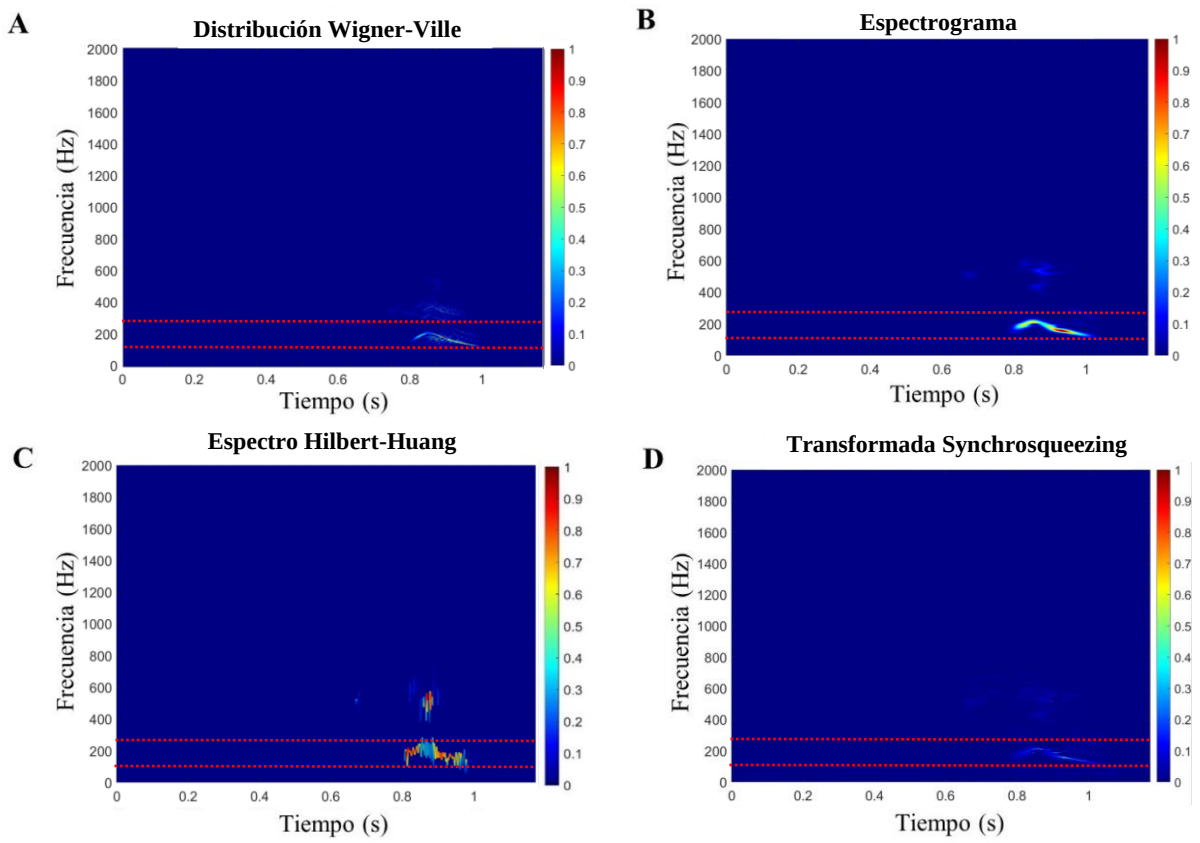

Fig. 28. TFRs de un ciclo de SR con sibilancia.

V.3.3. Ejemplo 2 de ciclo respiratorio con SR con sibilancia

Considerando otro ejemplo de un ciclo respiratorio que contiene una sibilancia polifónica, se presenta en la **Fig. 29A** la WVD de la señal, donde se aprecian componentes de alta energía en el tiempo 0.58 a 0.76 s, los cuales corresponden a una sibilancia. En los paneles de la **Fig. 29** la sibilancia se muestra entre líneas rojas punteadas que definen el rango espectral del sonido adventicio en cuestión. Cabe mencionar que en esta imagen aparecen tres componentes, en este caso particular los pertenecientes a sibilancias se atribuyen al componente superior y el inferior entre 100 y 600 Hz, mientras que el componente medio, alrededor de 350 Hz, se atribuye a un término cruzado, ya que no se observa en las representaciones obtenidas a partir de las otras tres TFR, i.e., SP, HHS y SST, para el mismo ciclo respiratorio. Mientras que la **Fig. 29B** muestra la TFR clásica de SP al emplear una ventana Blackman de 100 ms, y un traslape de ventana de 50%. En esta figura aparece una

sibilancia alrededor de 0.58 a 0.76 s, con un componente de alta energía con frecuencia de 200 Hz, el cual cumple con ambas características establecidas por CORSA [9] y otro componente de alta energía alrededor de 0.65 s y 600 Hz, sin embargo, debido a su duración no cumple con la característica temporal para ser considerado una sibilancia, pudiendo ser resonancia de la sibilancia. También se aplicó la técnica de HHS al mismo ciclo respiratorio, mostrado en la **Fig. 29C**, siguiendo los parámetros elegidos en la sección anterior: método de descomposición CEEMDAN y el uso de todos los IMF obtenidos a partir de ella, para no perder información importante que pueda ocasionar el ocultamiento de algún componente relevante para la detección posterior. En esta TFR aparece la sibilancia entre 0.58 y 0.76 s, alrededor de 150 Hz, como se visualiza también para las restantes TFR. Además, se presenta el componente de alta energía cerca de 0.65 s y 600 Hz. Esta TFR de alta resolución presenta los componentes más diferentes a nivel visual si lo comparamos con las otras representaciones, ya que la información de mayor energía se observa como componentes lineales verticales, en vez de presentarse como cúmulos pequeños y concentrados como en las otras TFR. En este sentido, la información de las sibilancias, como la presente en la **Fig. 29C** no aparece como una línea cuasi-horizontal, sino como un cúmulo de líneas verticales que no son continuas a lo largo del tiempo, por lo que a simple vista podría parecer no cumplir con la característica temporal de una sibilancia. En la **Fig. 29D** se presenta la SST al implementar una ventana Blackman de 100 ms. Esta TFR basada en el análisis de Fourier, comparte similitudes con el SP, de manera que se puede observar el parecido con esta representación, con la clara diferencia de la resolución. La SST presenta componentes mucho más finos y definidos tanto en el tiempo como en la frecuencia, es por esto por lo que es considerada como una TFR de alta resolución. En la **Fig. 29D** se muestra una sibilancia entre 0.58 y 0.76 s alrededor de los 150 Hz, y un componente de alta energía en 0.65 s cerca de los 600 Hz.

Con este ejemplo se busca resaltar la peculiaridad que presenta la WVD y la aparición de términos cruzados, ya que para el ojo no entrenado el término del medio podría pasar como una sibilancia, cuando no lo es realmente.

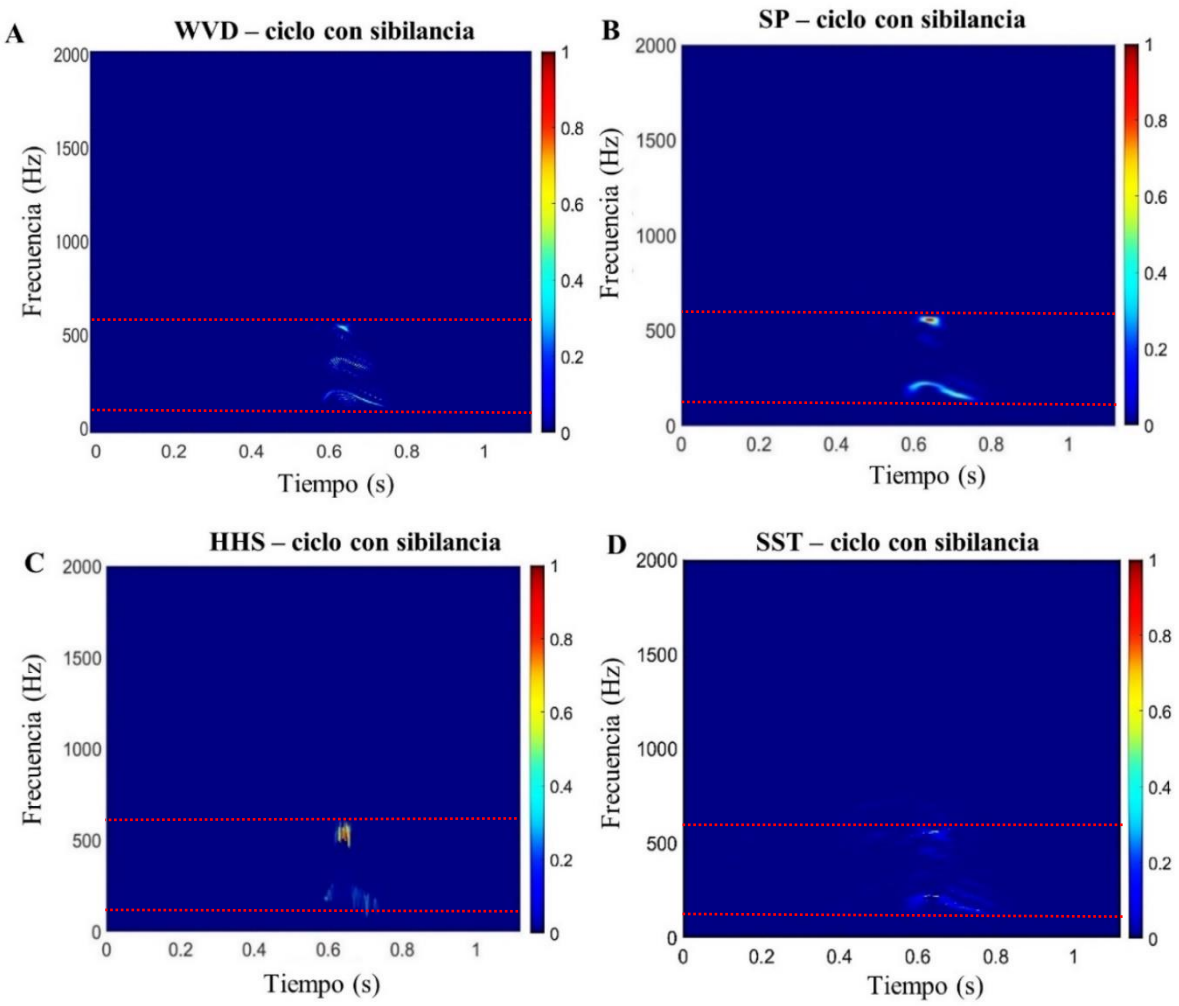

Fig. 29. Ejemplo de TFRs de ciclo respiratorio con sibilancias.

V.4. Detector de ciclos respiratorios normales y con sibilancias

En la etapa de pruebas preliminares se implementaron las estructuras de redes neuronales sencillas, mencionadas en la sección IV.5.1, que fueron entrenadas con un subconjunto de la mitad de los datos totales. Sin embargo, los resultados obtenidos con estas redes de pocas capas y que tenían que realizar todo el aprendizaje de los datos, obtuvieron un desempeño muy pobre, al ser de 50% de exactitud, como el mostrado en la **Fig. 30**, donde se observa un ejemplo de la gráfica de aprendizaje de la red al realizar un entrenamiento para un conjunto de imágenes de HHS, donde las probabilidades de detectar si un ciclo respiratorio es normal

o contiene sibilancias son comparables a lanzar una moneda. Así mismo, se exploró la implementación de redes propuestas anteriormente en otros trabajos, como la red desarrollada por Demir et al. [32], obteniendo resultados similares. Por lo anterior, se optó por buscar otra alternativa como la estrategia de transferencia de aprendizaje, aprovechando la ventaja de que las redes usadas para este método ya han sido entrenadas con un conjunto grande de imágenes y aprendieron características de ellas. Una de las técnicas adicionales para ampliar la cantidad de datos disponibles es el aumento de datos, sin embargo, en este trabajo de tesis no se exploró, por lo que se sugiere para emplear en un trabajo posterior.

Para realizar la detección y el proceso de entrenamiento, se empleó la aplicación *Deep Network Designer* del software MATLAB R2022b, y se dividieron los datos aleatoriamente para utilizar 70% de ellos para el conjunto de entrenamiento, 25% para el conjunto de validación y 5% para el conjunto de prueba. Es relevante destacar que, la mayoría de los trabajos encontrados en la literatura no emplea un grupo de prueba, solo utilizan conjuntos de entrenamiento y de validación. El conjunto de prueba empleado en este trabajo contiene 17 ciclos respiratorios normales y 17 con sibilancias pertenecientes a seis pacientes distintos,

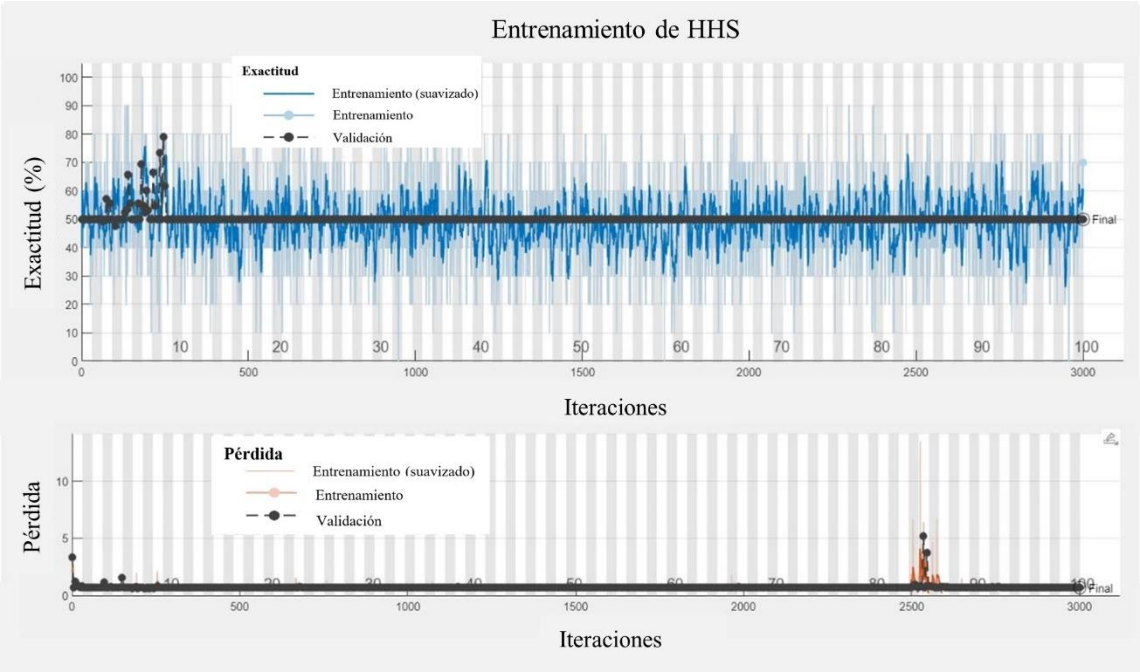

Fig. 30. Ejemplo de gráfica de aprendizaje en entrenamiento de HHS con red sencilla.

i.e., tres pacientes normales y tres con sibilancias, los cuales fueron seleccionados aleatoriamente de la base de datos ICBHI.

En la **Tabla VIII** se presentan los resultados de la búsqueda exhaustiva de hiperparámetros, mostrando aquellos que obtuvieron los mejores resultados de exactitud en entrenamiento y en validación, así como en pérdida de entrenamiento y de validación, para cada representación. Los parámetros de tamaño de lote y tasa de aprendizaje inicial produjeron resultados constantes para las cuatro TFR, obteniendo como resultados 64 y 0.0003 respectivamente, mientras que los tres parámetros restantes obtuvieron variación en los resultados.

A partir de los resultados de la búsqueda de hiperparámetros, se realizaron los entrenamientos empleando estos hiperparámetros. En la **Fig. 31** se muestra, como ejemplo, el resultado de la curva de aprendizaje obtenida en el proceso de entrenamiento y validación del SST. En el panel superior, en color azul claro, se muestra la tendencia de la exactitud para el entrenamiento, en azul oscuro la misma exactitud de entrenamiento suavizada, y en negro con línea punteada se muestra el resultado de la validación. Mientras que el panel inferior presenta la función de pérdida en color naranja y la misma con suavizado en color naranja oscuro, además de la validación en color negro con línea punteada.

El criterio de selección de paro para el entrenamiento fue el resultado de mejor pérdida de validación, o *best validation loss*, empleado como referencia para evitar el resultado erróneo presentado por la red en iteraciones posteriores, en las cuales se presenta sobreajuste a partir de la iteración 25, aproximadamente. Este criterio de paro se encuentra en las opciones de entrenamiento de la aplicación de MATLAB *Deep Network Designer*,

Tabla VIII. Resultados de búsqueda exhaustiva de hiperparámetros, para cada TFR

TFR	Épocas	Regularización L2	Frecuencia de validación
SP	50	1×10^{-5}	5
HHS	50	1×10^{-10}	30
SST	100	1×10^{-5}	5
WVD	150	1×10^{-5}	5

*Se obtuvo Tamaño de lote = 64 y Tasa de aprendizaje inicial = 0.0003 para todas las TFR

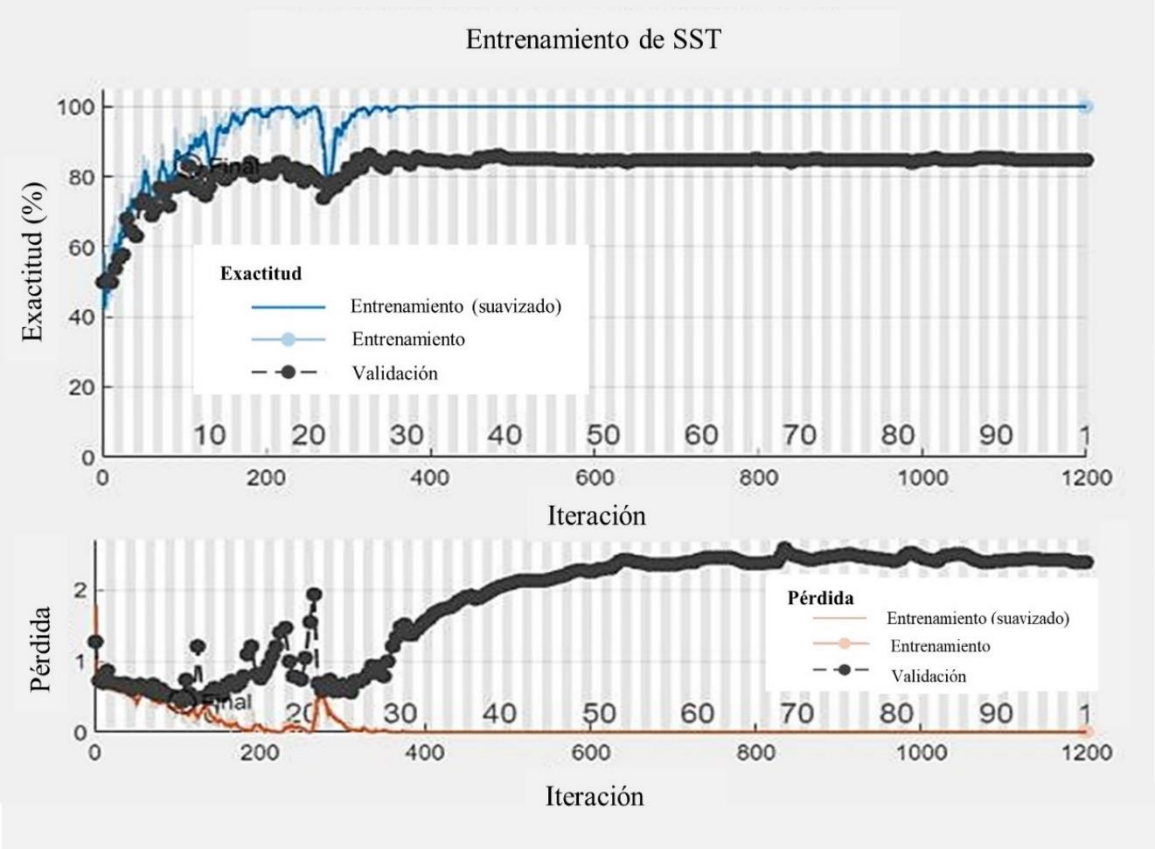

Fig. 31. Curva de aprendizaje en entrenamiento de SST.

donde se selecciona automáticamente el resultado de entrenamiento con mejor pérdida de validación. La precisión del entrenamiento fue del 84.14%, mientras que la precisión de la validación fue del 83.14% para SST.

Al realizar los entrenamientos se obtuvieron las matrices de confusión para los diferentes conjuntos de validación. Un ejemplo se muestra en la **Fig. 32**, donde se muestra la matriz de confusión obtenida con el SP para evaluar este conjunto, compuesto por 344 imágenes o TFRs. Este mismo procedimiento fue realizado para las cuatro TFR, para posteriormente calcular los índices propuestos anteriormente.

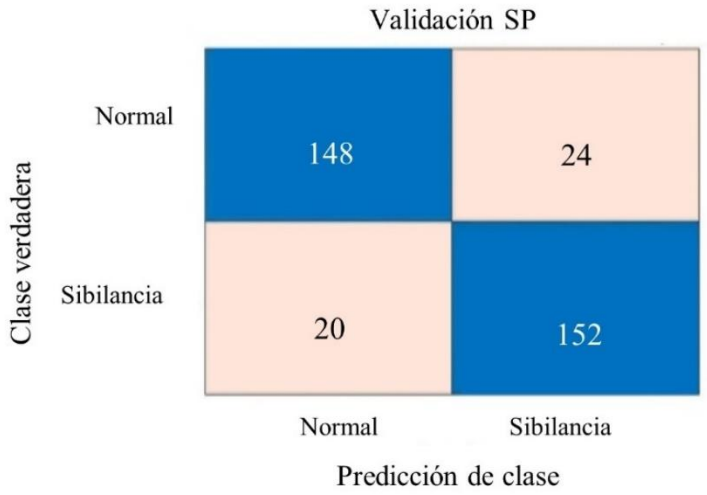

Fig. 32. Ejemplo de matriz de confusión de validación al evaluar la TFR de Espectrograma.

De acuerdo con los índices de desempeño a evaluar, la mejor exactitud en clasificación binaria para los datos de validación fue obtenida mediante la TFR del SP, la cual alcanzó una calificación de 0.872, superando a la SST por un rango muy pequeño, como se observa en la **Tabla IX**. De la misma manera, el SP consiguió el mejor resultado de especificidad con 0.883 y precisión con 0.880. Mientras que la TFR con mayor sensibilidad fue la SST con una puntuación de 0.912. De la misma manera, la SST obtuvo la mejor calificación de F_1-*Score* con 0.872. Mientras que el HHS reportó los menores resultados para los cuatro parámetros de clasificación. Los mejores resultados para cada índice se resaltan en negritas.

En la **Tabla X** se muestra el conjunto de datos empleado en este trabajo y cómo se dividieron para formar cada grupo a evaluar. Los datos comprenden 1184 ciclos respiratorios totales, de los cuales fueron retirados 34 para conformar el conjunto de prueba, donde 17 ciclos respiratorios pertenecen a tres sujetos que presentan sibilancias y los 17 restantes a tres sujetos sanos. De esta manera, 1150 ciclos se emplearon como datos de entrenamiento y validación. El conjunto de entrenamiento contiene 828 ciclos respiratorios, divididos entre 34 sujetos que presentan sibilancias y 33 sujetos normales. Mientras que el conjunto de validación se compone de 322 ciclos que pertenecen a 11 sujetos normales y 11 con sonidos adventicios continuos. Para realizar la verificación de desempeño de la red, se realizó

Tabla IX. Resultados obtenidos para el conjunto de validación de cada TFR

TFR	Exactitud	Sensibilidad	Especificidad	Precisión	F_1-Score
SP	**0.872**	0.860	**0.883**	**0.880**	0.870
HHS	0.779	0.790	0.767	0.772	0.781
SST	0.866	**0.912**	0.819	0.835	**0.872**
WVD	0.830	0.833	0.827	0.828	0.830

Tabla X. Conjuntos de datos para entrenamiento

Conjunto	Porcentaje	No. Sujetos	No. Ciclos respiratorios
Entrenamiento	70	34 Sibilancias 33 normales	828
Validación	25	11 Sibilancias 11 normales	322
Prueba	5	3 sibilancias 3 normales	34

validación cruzada con *k*-dobleces (*k-fold cross validation*) con $k = 5$ dobleces, al dividir aleatoriamente la base de datos, en cinco conjuntos de 115 imágenes de ciclos respiratorios normales y 115 imágenes de ciclos respiratorios con sibilancias, obteniendo finalmente 5 conjuntos de 230 imágenes, las cuales son distintas para cada conjunto y de distintos sujetos.

A partir de este análisis se obtuvieron los cálculos de la **Tabla XI**, donde se observan los resultados de exactitud obtenidos en cada una de las cinco iteraciones (*k*) y en el último renglón la media y desviación estándar (*standard deviation*, SD) de cada TFR. Una vez más, se observa que el SP es la TFR que obtuvo las calificaciones más altas de desempeño de detección de ciclos respiratorios con sibilancias, con un resultado de 84.60% ± 2.65 de exactitud. Sin embargo, la SST obtuvo una calificación de 84.34% ± 3.41 de exactitud, de manera que cualquiera de estas dos TFR puede ofrecer un buen índice de detección de sibilancias. Sin embargo, la SST es una técnica de alta resolución, lo que puede traer ventajas de visualización sobre el SP.

Tabla XI. Resultados para la exactitud obtenidos mediante validación cruzada para cada TFR

Iteración (k)	SP	HHS	SST	WVD
1	83.04	65.65	87.39	73.91
2	85.22	65.65	86.96	80.43
3	87.39	72.61	80.43	80.43
4	80.87	77.83	80.87	80.00
5	86.52	75.22	86.09	83.48
Media ± SD	84.60 ± 2.65	71.39 ±5.55	84.34 ± 3.41	79.65 ± 3.49

Mientras que, la WVD obtuvo un desempeño de 79.65% ± 3.49, lo cual puede deberse a la aparición de términos cruzados en la representación. Sin embargo, la información relevante perteneciente a la señal permanece en la TFR. A pesar de no ser una técnica empleada para la visualización o detección de sonidos adventicios continuos en los SR, la WVD ofrece buen desempeño para la clasificación de ciclos respiratorios con sibilancias, ya que obtuvo un resultado mayor a trabajos presentados en el estado del arte que emplean TFR comunes o de baja resolución [20], [30], [31], [32], [33]. Además, una de sus ventajas con respecto a ellas es que esta distribución es única para cada señal, por lo que no es necesario explorar o elegir parámetros que modifiquen la visualización de la WVD. Finalmente, el HHS obtuvo 71.39% ±5.55 de desempeño en la detección al emplear el conjunto de validación.

Además, se realizó un análisis estadístico de varianza de una vía para comparar los resultados de exactitud obtenidos mediante cada método TF, con un índice de confianza $\alpha = 0.05$ se encontró que existe una diferencia estadísticamente significativa entre las exactitudes de los clasificadores con las cuatro TFR, al obtener un valor *p* de 0.00019. Así mismo, se realizó otro análisis estadístico de varianza entre los grupos SP, SST y WVD, i.e., descartando el HHS que presentó el menor desempeño, para el mismo índice de confianza $\alpha = 0.05$ obteniendo un valor *p* de 0.133, indicando que no existe una diferencia estadística significativa entre estas tres últimas representaciones al evaluar la exactitud en el conjunto de validación.

Posteriormente, se emplearon las redes entrenadas con cada conjunto de imágenes TFR, para realizar la detección de ciclos respiratorios que contienen sibilancias usando el conjunto de prueba, donde finalmente se obtuvieron las matrices de confusión presentadas en la **Fig. 33**. Los resultados de este análisis se muestran en la **Tabla XII**. La TFR con mejores resultados es la SST, al obtener 0.705 en exactitud, 0.823 en sensibilidad, y empatar en 0.666 en precisión con el HHS, siendo que esta última TFR presentó el mejor índice de especificidad con 0.705. Cabe mencionar que ambas técnicas, SST y HHS, son las correspondientes al análisis TF de alta resolución, superando en este caso a la TFR de baja resolución, i.e., SP.

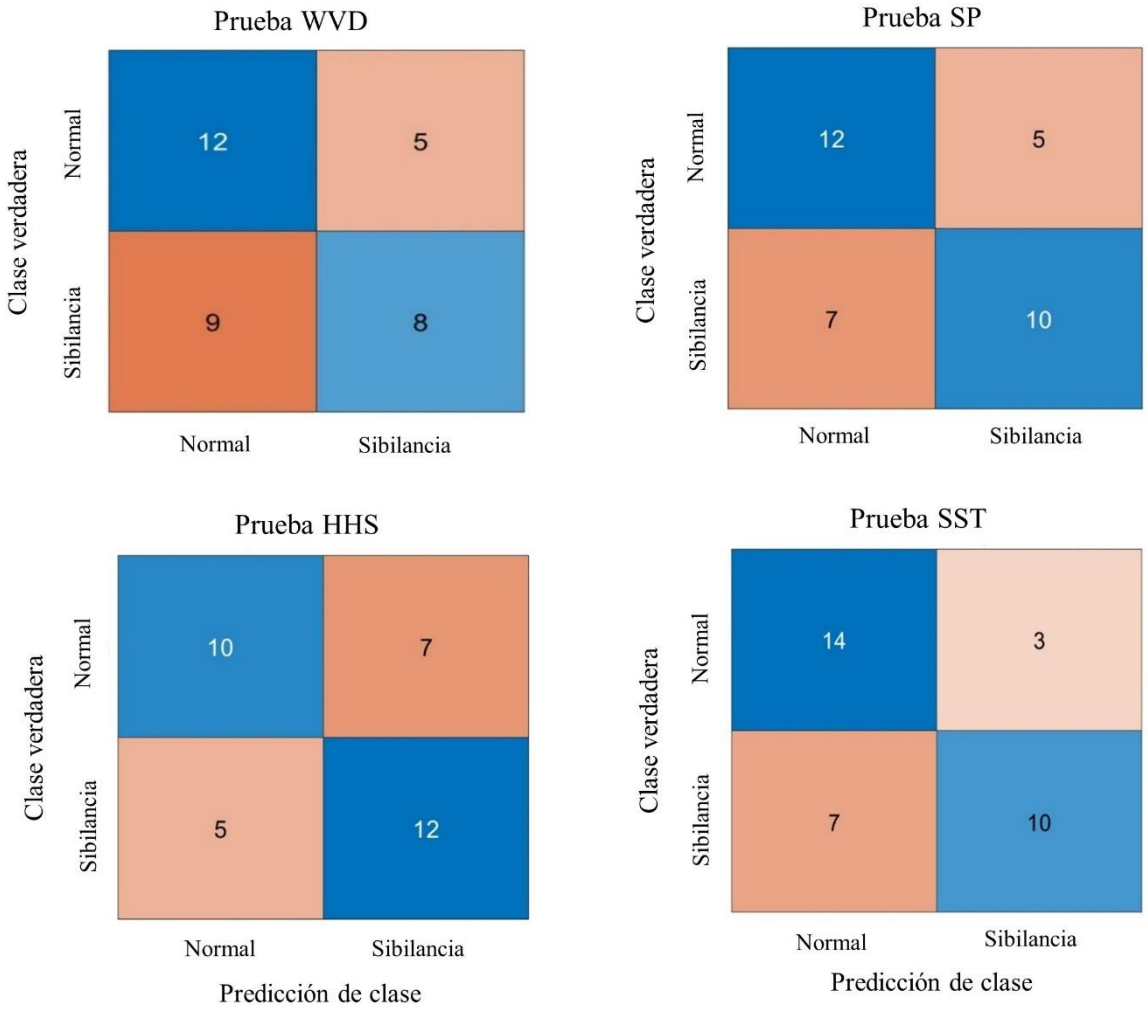

Fig. 33. Matrices de confusión de conjuntos de prueba obtenidas a partir de cada TFR. A) WVD. B) SP. C) HHS. D)SST.

Tabla XII. Resultados obtenidos para el conjunto de prueba de cada TFR

TFR	Exactitud	Sensibilidad	Especificidad	Precisión	F_1-Score
SP	0.647	0.705	0.588	0.631	0.666
HHS	0.647	0.588	**0.705**	**0.666**	0.624
SST	**0.705**	**0.823**	0.588	**0.666**	**0.736**
WVD	0.441	0.588	0.294	0.454	0.512

Con propósitos de comparación con otros trabajos, se realizó un cálculo de desempeño de validación general, obteniendo los resultados de la **Tabla XIII**, donde en el primer renglón se muestra el promedio de los índices calculados individualmente para cada TFR para el conjunto de validación, e.g., $Exactitud\ general = (Acc_{WVD} + Acc_{SP} + Acc_{HHS} + Acc_{SST})/4$. Así como el promedio de los índices calculados individualmente para el conjunto de prueba de cada TFR, en el segundo renglón, siguiendo el mismo principio, como una manera de evaluar el desempeño global del trabajo realizado. Como un método adicional para evitar el sobre-ajuste que se pueda presentar al realizar la clasificación se pudieran agregar capas de retiro o *dropout* o emplear técnicas de regularización L2 para reducir el valor de los parámetros, además de considerar cambiar la configuración de la red neuronal o incluso optar por emplear otros métodos de clasificación como los pertenecientes a aprendizaje de máquina.

Cabe mencionar que, los resultados obtenidos en este trabajo, al realizar la evaluación de desempeño con el conjunto de validación para cada TFR, por lo menos alcanzan a los descritos en la literatura donde emplean técnicas usuales para análisis de SR como los MFCC. En la **Tabla XIV** se muestran algunos trabajos que utilizan la misma base de datos ICBHI *Challenge* para realizar clasificación de los ciclos respiratorios. Sin embargo, los esfuerzos reportados difieren con este trabajo de tesis en el número de grupos a clasificar, ya que la mayoría de ellos realiza una clasificación en los cuatro grupos que contiene originalmente la base de datos ICBHI y solo uno de los aquí mencionados ejecuta una clasificación en tres grupos, el cual excluye al grupo combinado de ambos sonidos adventicios. Dentro de esas investigaciones se encuentra la realizada por Jakovljević et al. [30], donde reportan una sensibilidad de 56% o el resultado obtenido por Kochetov et al. [31] donde alcanzan 73%

con la misma técnica. Así como en el trabajo de Demir et al. [32], quienes lograron una exactitud de 71.1% al utilizar la técnica de SP y un clasificador de red neuronal convolucional, al emplear la misma base de datos, o el trabajo realizado por Ma et al. [22], quienes utilizan imágenes de SP y una red neuronal *ResNet*, obteniendo una sensibilidad de 41.3% y especificidad de 63.2%. Incluso en trabajos como el realizado por Shethwala et al.

Tabla XIII. Promedio de índices de las cuatro TFR, para los conjuntos de validación y prueba

Promedio TFRs	Exactitud	Sensibilidad	Especificidad	Precisión	F_1-Score
Validación	0.836 ± 0.042	0.848 ± 0.051	0.824 ± 0.047	0.828 ± 0.044	0.838 ± 0.042
Prueba	0.610 ± 0.115	0.676 ± 0.112	0.543 ± 0.175	0.604 ± 0.101	0.634 ± 0.093

[20], donde emplean transferencia de aprendizaje, logran una precisión de 64% y exactitud de 72% al utilizar espectrograma de Mel.

La excepción es el resultado obtenido por Chen et al. [23], quienes obtuvieron un desempeño con sensibilidad de 96.27%, especificidad de 100% y exactitud de 98.76% al emplear la transformada S optimizada y una red neuronal profunda para llevar a cabo una clasificación en tres grupos (1: sonidos normales, 2: crepitancias, y 3: sibilancias). Es importante mencionar que, la clasificación realizada en este trabajo de tesis solo comprende dos clases, en comparación a los distintos trabajos mencionados anteriormente, los cuales realizan la clasificación de la base ICBHI en las cuatro categorías que contiene originalmente la base de datos, o en tres categorías para el caso de Chen et al. [23], por lo que no son directamente comparables con los resultados obtenidos en este trabajo de tesis.

Tabla XIV. Estado-del-arte donde emplean la base de datos ICBHI Challenge

Título del trabajo	Año	Técnica utilizada	Clasificación	Resultado
N. Jakovljević et al. [30]	2018	Coeficientes cepstrales de Frecuencia de Mel + Modelo oculto de Markov	4 clases (N, C, S, C+S)	Se: 42.0 % Sp: 56.0 % *Score*: 49.5%
Kochetov et al. [31]	2018	Coeficientes cepstrales de Frecuencia de Mel → Red neuronal recurrente de enmascaramiento de ruido	4 clases (N, C, S, C+S)	Se: 58.4 % Sp: 73.0 % *Score*: 65.7 %
Chen et al. [23]	2019	Coeficientes de Transformada S optimizada → Espectrograma → DNN (ResNet)	3 clases (N, C, S)	Se: 96.2 % Sp: 100 % Acc: 98.7 %
Demir et al. [32]	2020	Espectrograma → CNN pre-entrenada + Análisis LDA + Método RSE	4 clases (N, C, S, C+S)	Se: 61.0 % Acc: 71.1% Prec: 69.0 %
Ma et al. [22]	2020	Espectrograma → DNN (ResNet modificada)	4 clases (N, C, S, C+S)	Se: 41.3 % Sp: 63.2 % *Score*: 52.2 %
Shethwala et al. [20]	2021	Espectrograma de Mel → CNN pre-entrenadas	4 clases (N, C, S, C+S)	Acc: 72.0 % Prec: 64.0 %

N: SR normal, C: SR con crepitancias, S: SR con sibilancias, C+S: SR con crepitancias y sibilancias.
CNN: red neuronal convolucional, DNN: red neuronal profunda, LDA: análisis discriminante lineal, RSE: método de conjuntos de subespacio aleatorio.
Se: sensibilidad (*sensitivity*), Sp: especificidad (*specificity*) Acc: exactitud (*accuracy*), Prec: precisión (*precision*), *Score*: (sensibilidad + especificidad)/2.

Capítulo 6

Conclusiones

En este trabajo, se propuso analizar algunas TFR de alta resolución TF conjunta que, en la literatura reportada previamente, han sido utilizadas en análisis de otro tipo de señales fisiológicas o que no han sido exploradas a profundidad para realizar clasificación automática de sibilancias. La técnica clásica empleada comúnmente para la detección de sibilancias, el SP, fue empleada en este trabajo de tesis como punto de comparación con las otras técnicas TF.

De acuerdo con los resultados, estas técnicas han logrado, por lo menos, alcanzar a aquellas que se utilizan comúnmente, e.g., MFCC o espectrograma de Mel, para la clasificación de sibilancias y eventos respiratorios, alcanzando puntuaciones equiparables en los índices de desempeño. Al realizar la validación cruzada con *k*-dobleces (*k*=5) se obtuvieron valores de exactitud de 84.60 % ± 2.65 para el SP, 84.34 % ± 3.41 para SST, 79.65 % ± 3.49 para WVD y 71.39% ±5.55 para HHS. Estos resultados de clasificación no presentan diferencias estadísticamente significativas, por lo que pueden ser consideradas para intentar resolver el problema de clasificación de sonidos respiratorios normales y con sibilancias. Interesantemente, a pesar de que una TFR de alta resolución ofrece una representación con mayor detalle e información precisa de una señal tanto en contenido en el tiempo como en la frecuencia, no necesariamente implica el desempeño más alto en exactiud en la tarea de clasificación, quizás por las características de las imágenes TF que produce cada representación. De esta manera, el SP, una técnica TF clásica, ofrece una imagen con cierto nivel de compromiso entre la resolución en tiempo y la resolución en frecuencia, que dependen de los parámetros a utilizar, siendo que el SP obtuvo buenos resultados en los índices de desempeño calculados en este trabajo, mientras que la técnica de alta resolución

SST obtuvo valores ligeramente menores, así como el HHS, el cual obtuvo los índices de desempeño más bajos de las cuatro TFR exploradas en este trabajo, al evaluar el conjunto de validación. Además, la WVD obtuvo valores equiparables a la SST y SP, por lo que se plantea continuar explorando esta representación como método auxiliar en la detección de sibilancias.

Además de calificaciones globales de exactitud de 83.6 %, sensibilidad de 84.8%, 82.4 % de especificidad y 82.8 % en el sistema de detección de ciclos respiratorios normales y con sibilancias, al realizar el entrenamiento con el conjunto de validación. Estos resultados alcanzan y sobrepasan a la mayoría de aquellos encontrados en la literatura como aquellos donde emplean técnicas como los MFCC y reportan una sensibilidad entre 56% y 73 %, así como al emplear el SP logrando una sensibilidad y exactitud de 40%. Incluso en trabajos donde emplean transferencia de aprendizaje, logran una precisión de 64% y exactitud de 72% al utilizar espectrograma de Mel. Sin embargo, hay excepciones como el trabajo realizado mediante transformada S optimizada donde logran sensibilidad de 96.27%, especificidad de 100% y exactitud de 98.76%. Cabe mencionar que los trabajos referidos realizaron una clasificación multiclase empleando tres o cuatro categorías pertenecientes a la base de datos empleada, mientras que en el presente trabajo se realizó una clasificación binaria utilizando solo la clase normal y clase con sibilancias, lo cual puede ser un factor para la disparidad en los resultados alcanzados y los encontrados en el estado del arte. Debido a la diferencia en la cantidad de grupos a evaluar de nuestro trabajo y de los esfuerzos reportados en la literatura, consideramos que los resultados de clasificación obtenidos no son directamente comparables entre sí. Es necesario continuar explorando estos métodos TF y aplicarlos en la totalidad de datos de la base para realizar una clasificación multiclase que permita una comparación adecuada con los trabajos reportados en el estado del arte, que permita realizar un análisis entre los distintos métodos TF empleados, así como el método, modelo o arquitectura de los clasificadores.

Consideramos que el uso de las técnicas TF exploradas en este trabajo de tesis pueden ayudar en la detección automática de sibilancias, para posteriormente apoyar en el diagnóstico de enfermedades obstructivas como asma y EPOC. Sin embargo, aún queda trabajo por explorar. A continuación, se comentan las áreas de oportunidad. Con respecto al

HHS, se requiere trabajar en el procesamiento posterior de los datos, implementando un suavizado para las imágenes de HHS, ya que fue la TFR que obtuvo menores resultados en los índices de desempeño resultantes en este trabajo. Debido a que el HHS es una representación de alta resolución, pudiera perder detalles importantes al entrar a la red neuronal para su clasificación. Así como, características propias de la descomposición o la construcción de esta representación que se han reportado previamente. Consideramos que se requieren diversas investigaciones con diferentes conjuntos de datos, no necesariamente sonidos respiratorios. Una de las bondades de las TFR es que son aplicables para una gran variedad de señales, en particular señales fisiológicas, de manera que aplicar estos métodos TF permite recopilar información que no muestra la señal a simple vista. Trabajar con diversos conjuntos de datos es un paso crucial para comprender la relación entre el desempeño de los clasificadores basados en representaciones tiempo-frecuencia, de resoluciones altas y bajas, con la resolución tiempo-frecuencia conjunta que estas ofrecen.

El método de clasificación propuesto se basó en la cantidad de datos a partir de la base de datos de acceso público ICBHI *Challenge* 2017. Se realizaron pruebas implementando redes neuronales sencillas para aprender características de la base de datos desde cero, sin embargo, los mejores resultados fueron obtenidos mediante transferencia de aprendizaje con la red pre-entrenada *GoogleNet*. Esta red fue modificada en sus últimas capas para ajustarse a los datos y, creemos que realizar otras transformaciones a estas capas y explorar otras arquitecturas o tipos de redes podría resultar en una mejor clasificación para esta pequeña base de datos u otras que cuenten con datos limitados.

Una de las limitaciones de este trabajo fue el relacionado a la adquisición de datos y el empleo de una base de datos pública. Debido al problema sanitario de COVID-19 no fue posible realizar adquisición de SR de pacientes sanos y con enfermedades pulmonares obstructivas, por lo que se optó por emplear una base de datos de acceso público. Al contar con más registros de sonidos se pueden vencer dos obstáculos: 1) la particularidad de la muestra de población a la que se va a evaluar, ya que se puede ampliar y complementar la base de datos con distintos grupos étnicos, de edades diversas y otras características, y 2) la dificultad que presenta entrenar una red neuronal de nueva creación con una cantidad pequeña de datos, ya que rápidamente puede caer en el sobreajuste de los datos.

Consideramos como una característica a resaltar la exploración de la WVD, la cual, al presentar artefactos como los términos cruzados, ha sido relegada y no se emplea para la clasificación de sibilancias. Esta distribución TF resulta complicada para el análisis visual humano, al ser difícil identificar los términos cruzados de los auto-términos. Sin embargo, al emplear un método computacional de detección, mediante inteligencia artificial, la red neuronal pudiera extraer las características de cada ciclo respiratorio, e.g., los intervalos o características propias de las sibilancias, y utilizarlas como información adicional para la predicción o clasificación de los eventos. Por otra parte, queremos exponer que, hasta nuestro conocimiento, la SST no ha sido empleada en la detección de sibilancias, por lo que comenzar a explorar las ventajas y desventajas que presenta este método para su aplicación en SR puede resultar motivador para trabajos futuros o incluso para otros investigadores y, de esta forma, continuar con la búsqueda de los parámetros que permitan adaptarse a una mayor variedad de datos o que obtengan mejores resultados de detección de eventos respiratorios o expandirse al análisis de otras señales.

Una característica importante de este trabajo es el empleo de un conjunto de prueba, para realizar clasificación y comprobar la eficiencia de la red. En otros trabajos encontrados en la literatura no se emplea un conjunto de prueba, o utilizan una parte de los datos de entrenamiento para probar la red. Sin embargo, realizar ese procedimiento resultará en un desempeño no determinante ya que son datos que la red ya había conocido al momento del entrenamiento. Nuestros resultados para los conjuntos de prueba son menores a los obtenidos en el conjunto de validación, donde la SST obtuvo la mejor exactitud con 70.5%, sensibilidad con 82.3%, precisión de 66.6% y F_1-*Score* de 73.6%; mientras que el HHS obtuvo el mejor resultado de especificidad con 70.5%. Sin embargo, es un resultado esperado al clasificar datos nuevos nunca vistos por la red durante el entrenamiento y validación cruzada. Además, es importante mencionar que los datos obtenidos de la base ICBHI presentan una gran variabilidad entre sujetos, al contar con un amplio rango de participantes: de todas las edades (desde infantes hasta adultos mayores), son personas de ambos sexos, distintas complexiones e índices de masa corporal y los sujetos presentan distintas patologías respiratorias con características diversas. Así mismo, los sonidos se registraron con diferentes dispositivos de adquisición y frecuencias de muestreo y presentan longitudes variables para cada ciclo respiratorio. Debido a estos factores, los resultados de clasificación con el conjunto de prueba

varían y emplear otro conjunto de datos de otros sujetos pudiera resultar en desempeños diferentes. No obstante, al evaluar el conjunto de validación se obtuvo un buen desempeño en los índices, por lo que consideramos que la clasificación con datos nunca vistos por la red resulta confiable.

Calcular la SST en el software MATLAB toma un tiempo ligeramente mayor al requerido para el SP, sin embargo, este esfuerzo se justifica al visualizar la SST y su alta resolución, que permite identificar los componentes de alta energía con mayor precisión tanto en el dominio temporal como en el espectral para la localización de sibilancias. Además de considerar que no hay diferencias estadísticamente significativas entre la SST, el SP y la WVD en términos de la exactitud de la clasificación, de manera que consideramos que la SST presenta el mejor desempeño en clasificación y en visualización de sibilancias. Así mismo, los resultados obtenidos mediante el conjunto de prueba presentan un sustento y justificación para la preferencia de esta técnica TFR sobre las demás, ya que este conjunto puede aproximarse al desempeño de la red en escenarios más realistas dentro de la práctica clínica.

Así mismo, planteamos que explorar la combinación de diversas técnicas TF presentadas en este trabajo pudiera lograr mejores resultados de clasificación que los mostrados en este proyecto de tesis, donde se realizaron individualmente, ya que se pueden combinar las bondades de cada una de las representaciones. Además, ampliar la búsqueda de parámetros para realizar cada una de ellas y encontrar una manera más eficiente de seleccionar estas variantes, como se demostró en secciones anteriores, tiene una gran influencia en la representación final, su forma y la información que contiene. Como otros métodos de búsqueda de hiperparámetros se plantea la exploración de otras técnicas como optimización bayesiana o búsqueda aleatoria, las cuales se pueden plantear para un trabajo futuro. Además, se pueden contemplar técnicas para aumentar los datos de la base ICBHI.

Finalmente, resulta interesante la posibilidad de ampliar la cantidad de datos para procesar la totalidad de registros de la base empleada y de esta manera explorar la detección multiclase utilizando las cuatro categorías que comprende la base de datos y ampliar el panorama para las técnicas que ofrece el aprendizaje profundo para incrementar el número de datos o balancear las clases. Así mismo, contar con una base más amplia de registros

realizados con distintos equipos de adquisición y protocolos diferentes resulta relevante para evaluar el desempeño del detector al incluir mayor variabilidad de criterios que posteriormente ofrezcan una mejor clasificación en SR de prueba.

Anexo

Artículo presentado en Congreso Internacional de IEEE *1st Conference on Medicine and Biology 2023 Region 9*

Assessment of an automated wheezing respiratory cycle identification based on CNN and classical and high-resolution time-frequency representations [56].

Assessment of an automated wheezing respiratory cycle identification based on CNN and classical and high-resolution time-frequency representations.

Amaranta P. Córdova-Martínez, *Student Member, IEEE*, Bersaín A. Reyes, *Member, IEEE*, Sonia Charleston-Villalobos and Tomás Aljama-Corrales

***Abstract*— Respiratory sounds (RS) provide relevant information that helps diagnose pathologies. Some common continuous adventitious RS are wheezes, present in diseases such as asthma and chronic obstructive pulmonary disease (COPD). In recent decades, efforts have been made to automatically detect wheezes in RS, e.g., by classical time-frequency (TF) representations such as the spectrogram, which allows the visualization of instantaneous changes in frequency over time. However, due to the characteristics and limitations of the previously used TF representations, the detection of wheezes with these techniques still offers areas of opportunity. In this work, we propose to perform the classification of normal RS cycles and RS cycles containing wheezes using two TF methods: (1) Spectrogram (SP) and (2) Hilbert-Huang Spectrum (HHS), as input to a pre-trained convolutional neural network (CNN), the GoogleNet. The proposed method was evaluated on the ICBHI Challenge 2017 database, using 575 wheezing respiratory cycles from 45 patients and 575 normal cycles from 44 subjects. The best experimental result achieved a specificity of 0.883, precision of 0.880, and accuracy of 0.872 for the validation set using SP method. Comparing this performance with those reported in the literature, it was found that the classification results obtained are at least comparable to those reported in most state-of-the-art works.**

***Clinical Relevance*— SP and HHS techniques allow the identification of wheezes and their temporal and spectral characteristics. Furthermore, the automatic classification of respiratory cycles, including wheezes, could help diagnose obstructive pulmonary diseases, obtaining better results than those found in state-of-the-art works, where standard TF methods are used, e.g., Mel spectrogram.**

I. Introduction

According to the World Health Organization, lung diseases are ranked among the ten leading causes of death worldwide. Respiratory pathologies such as asthma and chronic obstructive pulmonary disease (COPD) are non-communicable, chronic diseases that share symptoms such as narrowing of the airways, respiratory distress, and the appearance of wheezes in respiratory sounds (RS) produced by air turbulence through the airways. In healthy subjects, RS are soft and low-pitched noises. Any content that differs from these characteristics suggests abnormality or the presence of diseases, so the RS are relevant indicators of health and pathological conditions. RS are classified as normal and adventitious. The latter are considered to be added or superimposed on normal RS and can be classified as discontinuous, e.g., crackles, and continuous, e.g., wheezes, according to their duration [1].

One of the first diagnostic methods commonly used to detect respiratory diseases is the pulmonary auscultation. This is a fast, non-invasive method, which allows collecting evidence that is later confirmed by other more specialized techniques. However, this method has limitations, and hence in recent decades, computerized methods have been developed to detect adventitious breath sounds.

A straightforward technique involves visual inspection of peaks in the frequency domain, such as the work performed in 1989 by Pasterkamp et al., who compute the spectrogram (SP) of normal and adventitious RS. It is worth mentioning that an automatic classification of wheeze was not carried out; however, this work pioneered the visual information of the waveform and spectrum of the RS together with their joint time-frequency (TF) representation (TFR) [2].

Deep learning algorithms allow using an image, e.g., a TFR with time-varying spectral information about a signal, as input to the network, e.g., a convolutional neural network (CNN), to recognize patterns and classify data. CNN advantages include their architecture flexibility to improve performance, and the use of transfer learning strategies to deal with small databases.

Over the years, many techniques have been used for classification of normal RS and RS with wheezes, usually through neural networks, such as the work done by Demir et al., who performed classification using Mel spectrogram images as input to a CNN [3]. Of particular relevance are the techniques based on TF analysis, which has been shown to improve detection performance, compared to other techniques, by including both temporal and spectral criteria of wheezes in its TFR, allowing to detect its main time-varying spectral characteristics.

*Research supported by CONAHCYT, through a master's degree grant with CVU number: 1177680.

Amaranta P. Córdova-Martínez and Bersaín Alexander-Reyes are with the School of Sciences, Universidad Autónoma de San Luis Potosí, SLP, 78295, Mexico (a194230@alumnos.uaslp.mx, corresponding author phone: +52 444 823 2300 ext. 5676; e-mail: bersain.reyes@uaslp.mx).

Sonia Charleston-Villalobos and T. Aljama-Corrales are with the Electrical Engineering Department, Universidad Autónoma Metropolitana Iztapalapa, Mexico City 09340, Mexico (schv@xanum.uam.mx, alja@xanum.uam.mx).

The first works focused on TF analysis by SP to visualize wheezes. However, it is well known that the SP presents an inherent compromise between its temporal and spectral resolution due to the use of a fixed length window to produce the TFR. As an alternative to this method, high-resolution TF techniques have been developed. One of them is the Hilbert-Huang Spectrum (HHS), developed to analyze non-linear and non-stationary data, allowing to locate events in time, as well as in frequency from the decomposition of the signal under analysis into its intrinsic mode functions (IMF) using the Empirical Mode Decomposition (EMD) algorithm [4].

In this work, it is proposed to compare the automated classification of RS cycles containing, or not, wheezes using a classical and a high-resolution TFR. In particular, the SP and the HHS were explored to test if the latter, with a higher joint TF resolution, allows to obtain a better classification performance than the SP. Although it is expected that using a high-resolution TFR will allow a better description of wheeze energy in a concentrated manner in the joint time-frequency domain, it is not known, to the best of our knowledge, if this advantage will be reflected in the classification task. To this end, the performance of using each TFR method as input to a classifier was compared in a public database containing RS, as described below.

II. Background

A. Respiratory sounds

RS contain valuable information on physiology and pathology of the airways and the lungs. The main components of RS are found in a frequency band between 100 Hz and 2500 Hz [1]. RS reflect changes depending on the pathology in the lungs or in the airways, e.g., bronchial obstruction such as that seen in asthma increases high-frequency, high-intensity components. RS are typically classified as normal or adventitious. The resultant waveform of normal sounds is disorganized, i.e., contains many different frequencies. Adventitious RS can be classified as continuous or discontinuous, depending on their duration.

Within the continuous adventitious sounds are wheezes, characterized by having a longer duration than discontinuous adventitious sounds and are defined as a high-pitched sound with a duration greater than 100 ms and a dominant frequency component above 100 Hz [1].

B. Time-Frequency Analysis

To achieve greater performance in detecting wheezes, the classification models can include a set of joint criteria in the TF domain, pertaining to the duration, pitch range, and magnitude of the wheezes in the TFR, e.g., obtained via the classical SP [2]. These TFRs aim to use a function that describes the power of the RS simultaneously in the temporal and spectral domains to perform a complete analysis of their properties. The two TFR explored in this work, i.e., the SP and HHS, are briefly described next.

C. Spectrogram

Traditional Fourier analysis implies the idea that frequencies are not localized in time. Although this technique highlights the different tones involved in a signal, it is inefficient in describing their respective occurrences in time. However, the short-time Fourier transform (STFT) has been developed from this analysis leading to the TFR called spectrogram. The SP is defined as the squared magnitude of the STFT. That is, the corresponding energy density spectrum of the signal $s(t)$ at time t and frequency ω is given by:

$$\mathrm{SP}(t,\omega)=\left|\frac{1}{\sqrt{2\pi}}\int e^{-j\omega\tau}s(\tau)h(\tau-t)d\tau\right|^2 \quad (1)$$

where $h(\tau-t)$ is a window function centered at t. When multiplying by the window function $h(t)$, the original signal is emphasized at time t and suppressed at time τ. For each time instant, a different spectrum is obtained, and the totality of this spectrum is the TFR called SP.

One of the advantages of the SP is its ability to reintroduce the notion of time so that instantaneous changes in frequency are observed. It represents, in a two-dimensional plane, the energy of the signal against time and frequency [5]. However, this technique requires a selection of parameters that directly impact the frequency resolution of the TFR, such as the type and length of the window, as well as the overlap between adjacent segments.

D. Hilbert-Huang Spectrum

The HHS method considers the regional energy in the TF plane of each IMF of the signal, resulting from its EMD, as shown in (2). As a result of the sifting process produced by an adaptive decomposition, the signal $s(t)$ can be expressed as a linear combination of is components as:

$$\mathrm{s}(t)=\sum_{i=1}^{n} IMF_i(t)+r_n(t) \quad (2)$$

where n is the number of extracted IMFs and $r_n(t)$ is the residue of $s(t)$. By applying the Hilbert transform to each IMF, a complex signal is obtained, called the analytical signal, from which it is possible to estimate its instantaneous frequency and instantaneous amplitude as a function of time. The representation of the instantaneous frequencies and amplitudes of all the IMFs results in the representation called HHS, which offers a TFR of high joint TF resolution [4].

The HHS presents advantages when compared to traditional techniques such as Fourier analysis since the HHS is an adaptable technique and allows the adjustment of the method directly from the data, as well as considering the regional energy in the local TF plane and provides an advanced representation of the TF characteristics of multicomponent signals. In the RS field, the HHS has been used for wheeze detection [6]. However, due to its local nature, this method has a problem: oscillations with similar scales can be produced in different modes, called mode mixing (MM). To reduce this problem, methods such as the ensemble EMD (EEMD) have been proposed, which obtains more regular IMFs. However, this technique has created new difficulties, e.g., the reconstructed signal contains residual noise. The complete ensemble EMD with adaptive noise (CEEMDAN) method has solved this problem, reducing the reconstruction error to a non-significant level.

III. Methodology

A. Database

A portion of the database of RS signals belonging to the 2017 Challenge of the International Conference on Biomedical and Health Informatics (ICBHI Challenge) was employed.

Table I shows the selection by groups made in the ICBHI Challenge database, where the first group includes 575 respiratory cycles marked by the experts as cycles that contain wheezes. These cycles are found in 129 audio records belonging to 45 patients. For the group of normal subjects, the same number of normal respiratory cycles were randomly selected, which are contained in 77 audio records belonging to 44 different subjects. The previous strategy was used to perform balanced training both in the number of respiratory cycles and in the number of patients.

B. Pre-processing

The signals in the database have a sampling frequency of 4 kHz, 10 kHz, or 44.1 kHz. So, a single sampling frequency of 4 kHz was established to standardize subsequent processing, complying with the Nyquist criteria for RS. In addition, a high-pass FIR filter with a cut-off frequency of 100 Hz was applied, selected according to the wheeze characteristics reported in the literature [1], as well as to minimize the presence of heart sounds.

C. Simulation of synthetic signals and selection of TFR parameters

For each TFR, its parameters were selected by simulating a synthetic signal that complies with the temporal characteristics of a wheeze, i.e., duration greater than 100 ms, and spectral characteristics, i.e., containing a principal frequency component greater than 100 Hz. The synthetic signal was simulated as a polyphonic continuous adventitious sound with sinusoidally and linearly modulated frequencies, following the work of Lozano et al. [7]. The instantaneous frequency was calculated from the generated signal, obtaining its theoretical ideal TFR. Subsequently, a qualitative comparison was made between TF images obtained using different parameters for SP and HHS, with their theoretical ideal TFR. The comparison was made using various window lengths and types for the SP, and various EMD methods for the HHS. Finally, the 100 ms Blackman window was selected as the window that best approximates the simulated ideal TFR for the SP. The CEEMDAN method was selected for the HHS.

D. Classifier

The transfer learning strategy was implemented using the *GoogleNet* pre-trained network, whose final, or output, layers were modified to adapt it to train with the respiratory data. Fig. 1 illustrates the proposed neural network architecture with an SP image for a wheezing respiratory cycle as input for RS classification. From layer 140, the layers shown in the diagram were added to the pre-trained network. Likewise, the model was regularized to avoid overfitting the network by adding a 50% dropout layer. In addition, several modifications were made to the last layers of the network to debug it, and its hyperparameters were adjusted, as mentioned below.

TABLE I. DETAILS OF THE NORMAL AND WHEEZES GROUP EXTRACTED FROM THE ICBHI DATABASE.

Group	Number of audio records	Number of patients	Number of respiratory cycles
Wheezes	129	45	575
Normal	77	44	575

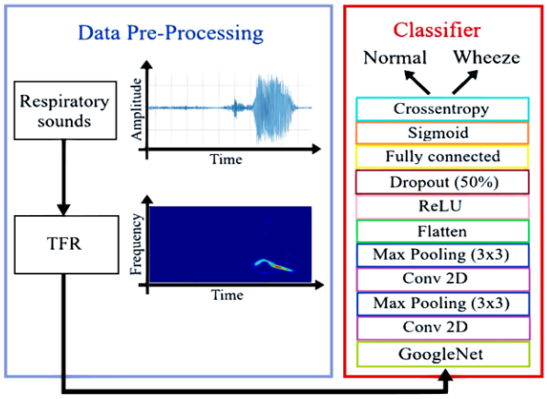

Figure 1. Overview of the proposed method.

The input to the network consists of TFR images obtained using the two proposed techniques. The TFR images were resized to reach a suitable size for the network, 224 x 224 pixels, where each image corresponds to a labeled respiratory cycle, using 70% of the data for the training set and 30% for the validation set. The network output comprises a sigmoid function, classifying into two categories, i.e., normal RS or wheeze-containing RS.

An exhaustive search of hyperparameters was conducted to select those with better accuracy in the classification for each TFR, with different options for the parameters such as number of epochs, batch size, L2 regularization, validation frequency, and initial learning rate.

E. Classifier performance analysis

The performance of the automatic wheeze identifier was based on information from the event annotations in the ICBHI database made by experts. The quantitative performance criteria were accuracy, sensitivity, specificity, precision, and F1-score.

IV. RESULTS

After an exhaustive search, the parameters resulting in the best accuracy results were as follows. For both, the SP and HHS TFRs, the number of epochs was 50, batch size was 64, and initial learning rate was 0.0003, while the L2 regularization was 1 x 10^{-5} and 1 x 10^{-10}, and for the validation frequency were 5 and 30, for the SP and HHS TFRs, respectively.

Fig. 2 shows the results of the learning curves obtained for the training and validation sets for the HHS. The selection

TABLE II. RESULTS OBTAINED FOR THE VALIDATION SET OF EACH TFR.

TFR	Accuracy	Sensitivity	Specificity	Precision	F1-Score
SP	0.872	0.860	0.883	0.880	0.870
HHS	0.779	0.790	0.767	0.772	0.781

criterion for training was the result with the best validation loss, as a technique to reduce network overfitting. Training accuracy was 77.8%, and validation accuracy was 71.6%.

When performing the training for each TFR, the confusion matrices for the validation set were obtained. An example is shown in Fig. 3, where the first panel, Fig. 3A, shows the confusion matrix obtained by evaluating the SP validation set, which includes 172 SP images, while the second panel, Fig. 3B, presents the one obtained for the HHS validation set.

According to the indices to be evaluated, the SP obtained the best performance for the five indices in the classification, reaching an accuracy of 0.872, surpassing the HHS, as observed in Table II. In the same way, the SP achieved the best sensitivity result of 0.860, specificity of 0.883, precision of 0.880 and F1-score of 0.870. In contrast, HHS got results between 0.767 and 0.790.

It is worth mentioning that the results in this work exceed some described in the literature, such as the work of Demir et al. [3], who achieved a sensitivity and accuracy of 40% when using the SP technique and a CNN classifier using the same database. Even in works such as the one performed by Shethwala et al. [8], where comparing different transfer learning techniques, they achieved a precision of 64% and an accuracy of 72% when using a Mel spectrogram. However, the classification task in our study was a binary one.

Using the ICBHI database represented a challenge due to the relatively low amount of data to train a neural network, but the use of transfer learning allowed RS classification with low- and high- resolution TFRs. Another challenge is the available labeling of the data, which sometimes seems controversial, and this could be reflected in low classification performance.

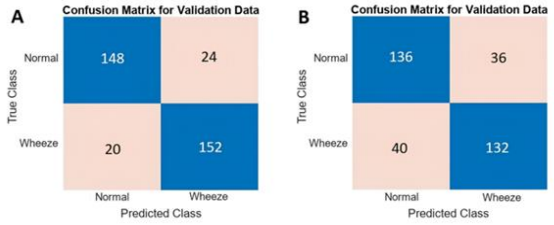

Figure 3. Confusion matrices obtained by classifying the validation set of: A) SP and B) HHS.

Figure 2. Training, validation, and loss deviation for HHS

V. Conclusion

According to the results, the techniques explored in this work have managed to surpass those that are commonly used, e.g., Mel frequency cepstral coefficients (MFCC) or Mel spectrogram, for the detection of wheezes and respiratory events, reaching higher scores of sensitivity, specificity, precision, and accuracy. Therefore, we consider that using these techniques can help to detect wheezes, to support the diagnosis of obstructive diseases such as asthma and COPD. However, there is still work to be explored regarding the subsequent processing of the HHS given the results reported in this work.

The SP provided the most high-performance results for the validation set, even though it is not a high-resolution time-frequency technique. This result appears to indicate that although high-resolution techniques improve TF analysis and readability, by having more defined components in time and frequency at a visual level, they will not necessarily obtain the best results for classification tasks. The HHS obtained the second-best classification results, still surpassing those presented in the literature.

The proposed classification method was selected based on the amount of data used from the ICBHI database. There is still work to be done, such as processing all the audio records contained in the database to perform training and data classification. The classification performed in this work was binary, by only using normal and wheezing registers, in comparison with other works found in state-of-the-art, where they implement 3 or 4 classes.

We consider worthy to explore the combination of different TF techniques to test if better classification results than those shown in this work, where they were employed individually, can be achieved.

References

[1] A. Sovijärvi, «Characteristic of breath sounds and adventitious respiratory sounds», n.º 20, pp. 591-596, 2000.

[2] H. Pasterkamp, C. Carson, D. Daten, y Y. Oh, «Digital Respirosonography», *Chest*, vol. 96, n.º 6, pp. 1405-1412, dic. 1989, doi: 10.1378/chest.96.6.1405.

[3] F. Demir, A. M. Ismael, y A. Sengur, «Classification of Lung Sounds With CNN Model Using Parallel Pooling Structure», *IEEE Access*, vol. 8, pp. 105376-105383, 2020, doi: 10.1109/ACCESS.2020.3000111.

[4] N. E. Huang *et al.*, «The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis», *Proc. R. Soc. Lond. A*, vol. 454, n.º 1971, pp. 903-995, mar. 1998, doi: 10.1098/rspa.1998.0193.

[5] L. P. Malmberg *et al.*, «Basic techniques for respiratory sound analysis», *European Respiratory Review*, 2000.

[6] Ö. Şayli, «Wheeze detection in the respiratory sounds using Hilbert-Huang transform.», *22nd Signal Processing and Communications Applications Conference (SIU)*, 2014, doi: 10.1109/SIU.2014.6830699.

[7] M. Lozano, J. A. Fiz, y R. Jané, «Performance evaluation of the Hilbert–Huang transform for respiratory sound analysis and its application to continucus adventitious sound characterization», *Signal Processing*, vol. 120, pp. 99-116, mar. 2016, doi: 10.1016/j.sigpro.2015.09.005.

[8] R. Shethwala, S. Pathar, T. Patel, y P. Barot, «Transfer Learning Aided Classification of Lung Sounds-Wheezes and Crackles», en *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India: IEEE, abr. 2021, pp. 1260-1266. doi: 10.1109/ICCMC51019.2021.9418310.

Referencias

[1] World Health Organization, «Las 10 principales causas de defunción», WHO, dic. 2020. [En línea]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] World Health Organization, «Newsroom: Fact sheet on Chronic obstructive pulmonary disease (COPD)», WHO, may 2022. [En línea]. Disponible en: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

[3] F. Cano Valle, *Enfermedades del aparato respiratorio*, 3.ª ed. México: Méndez Editores, 2013.

[4] World Health Organization, «Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019.», Geneva, 2020.

[5] J. E. Hall y A. C. Guyton, *Guyton y Hall, Tratado de fisiología médica*, 12.ª ed. España: Elsevier, 2011.

[6] W. Cristancho Gómez, *Fundamentos de fisioterapia respiratoria y ventilación mecánica*, 3.ª ed. Manual moderno, 2015.

[7] Y.-F. Li, B. Langholz, M. T. Salam, y F. D. Gilliland, «Maternal and Grandmaternal Smoking Patterns Are Associated With Early Childhood Asthma».

[8] G. Viegi. *et al.*, «Indoor air pollution and airway disease.», *International journal of tuberculosis and lung disease*, vol. 8, n.º 12, pp. 1401-1415, 2004.

[9] A. Sovijärvi, «Characteristic of breath sounds and adventitious respiratory sounds», n.º 20, pp. 591-596, 2000.

[10] N. Gavriely, Y. Palti, y G. Alroy, «Spectral characteristics of normal breath sounds», *Journal of Applied Physiology*, vol. 50, n.º 2, pp. 307-314, feb. 1981, doi: 10.1152/jappl.1981.50.2.307.

[11] A. R. A. Sovijärvi, J. Vanderschoot, y J. E. Earis, «Standardization of computerized respiratory sound analysis», *European Respiratory Review*, p. 65, 2000.

[12] N. Gavriely, «Breath sounds methodology», en *Breath Sounds Methodology*, 1st ed., New York: CRC Press, 1995, p. 203.

[13] H. Pasterkamp, C. Carson, D. Daten, y Y. Oh, «Digital Respirosonography», *Chest*, vol. 96, n.º 6, pp. 1405-1412, dic. 1989, doi: 10.1378/chest.96.6.1405.

[14] Q. Zhang *et al.*, «Noise Removal of Tracheal Sound Recorded During CPET to Determine Respiratory Rate», en *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany: IEEE, jul. 2019, pp. 4650-4653. doi: 10.1109/EMBC.2019.8857738.

[15] B.-S. Lin, H.-D. Wu, y S.-J. Chen, «Automatic Wheezing Detection Based on Signal Processing of Spectrogram and Back-Propagation Neural Network», *Journal of Healthcare Engineering*, vol. 6, n.º 4, pp. 649-672, dic. 2015, doi: 10.1260/2040-2295.6.4.649.

[16] L. J. Hadjileontiadis y S. M. Panas, «Nonlinear analysis of musical lung sounds using the bicoherence index», en *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. «Magnificent Milestones and Emerging Opportunities in Medical Engineering» (Cat. No.97CH36136)*, Chicago, IL, USA: IEEE, 1997, pp. 1126-1129. doi: 10.1109/IEMBS.1997.756551.

[17] J. A. Fiz *et al.*, «Analysis of Forced Wheezes in Asthma Patients», *Respiration*, vol. 73, n.º 1, pp. 55-60, 2006, doi: 10.1159/000087690.

[18] J. A. Fiz, R. Jané, A. Homs, J. Izquierdo, M. A. García, y J. Morera, «Detection of Wheezing During Maximal Forced Exhalation in Patients With Obstructed Airways», *Chest*, vol. 122, n.º 1, pp. 186-191, jul. 2002, doi: 10.1378/chest.122.1.186.

[19] M. Bahoura y C. Pelletier, «New parameters for respiratory sound classification», en *CCECE 2003 - Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No.03CH37436)*, Montreal, Que., Canada: IEEE, 2003, pp. 1457-1460. doi: 10.1109/CCECE.2003.1226178.

[20] R. Shethwala, S. Pathar, T. Patel, y P. Barot, «Transfer Learning Aided Classification of Lung Sounds-Wheezes and Crackles», en *2021 5th International Conference on Computing*

Methodologies and Communication (ICCMC), Erode, India: IEEE, abr. 2021, pp. 1260-1266. doi: 10.1109/ICCMC51019.2021.9418310.

[21] B. M. Rocha *et al.*, «An open access database for the evaluation of respiratory sound classification algorithms», *Physiol. Meas.*, vol. 40, n.º 3, p. 035001, mar. 2019, doi: 10.1088/1361-6579/ab03ea.

[22] Y. Ma, X. Xu, y Y. Li, «LungRN+NL: An Improved Adventitious Lung Sound Classification Using Non-Local Block ResNet Neural Network with Mixup Data Augmentation», en *Interspeech 2020*, ISCA, oct. 2020, pp. 2902-2906. doi: 10.21437/Interspeech.2020-2487.

[23] H. Chen, X. Yuan, Z. Pei, M. Li, y J. Li, «Triple-Classification of Respiratory Sounds Using Optimized S-Transform and Deep Residual Networks», *IEEE Access*, vol. 7, pp. 32845-32852, 2019, doi: 10.1109/ACCESS.2019.2903859.

[24] A. S. Lyons y R. J. Petrucelli, *La historia de la medicina*. Barcelona, España: Doyma, 1984.

[25] R. L. Murphy, S. K. Holford, y W. C. Knowler, «Visual lung-sound characterization by time-expanded waveform analysis», *The New England journal of medicine*, vol. 296, n.º 17, pp. 968-971, 28 de abril de 1977.

[26] L. P. Malmberg *et al.*, «Basic techniques for respiratory sound analysis», *European Respiratory Review*, 2000.

[27] Y. Shabtai-Musih, J. B. Grotberg, y N. Gavriely, «Spectral content of forced expiratory wheezes during air, He, and SF6 breathing in normal humans», *Journal of Applied Physiology*, vol. 72, n.º 2, pp. 629-635, feb. 1992, doi: 10.1152/jappl.1992.72.2.629.

[28] S. A. Taplidou *et al.*, «On Applying Continuous Wavelet Transform in Wheeze Analysis», en *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, CA, USA: IEEE, 2004, pp. 3832-3835. doi: 10.1109/IEMBS.2004.1404073.

[29] S. A. Taplidou, L. J. Hadjileontiadis, T. Penzel, V. Gross, y S. M. Panas, «WED: An efficient wheezing-episode detector based on breath sounds spectrogram analysis», en *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*, Cancun, Mexico: IEEE, 2003, pp. 2531-2534. doi: 10.1109/IEMBS.2003.1280431.

[30] N. Jakovljević y T. Lončar-Turukalo, «Hidden Markov Model Based Respiratory Sound Classification», *Precision Medicine Powered by pHealth and Connected Health*, vol. 66, pp. 39-43, nov. 2017.

[31] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, y A. Shalyto, «Noise Masking Recurrent Neural Network for Respiratory Sound Classification», en *Artificial Neural Networks and Machine Learning – ICANN 2018*, vol. 11141, V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, y I. Maglogiannis, Eds., en Lecture Notes in Computer Science, vol. 11141. , Cham: Springer International Publishing, 2018, pp. 208-217. doi: 10.1007/978-3-030-01424-7_21.

[32] F. Demir, A. M. Ismael, y A. Sengur, «Classification of Lung Sounds With CNN Model Using Parallel Pooling Structure», *IEEE Access*, vol. 8, pp. 105376-105383, 2020, doi: 10.1109/ACCESS.2020.3000111.

[33] S. Gairola, F. Tom, N. Kwatra, y M. Jain, «RespireNet: A Deep Neural Network for Accurately Detecting Abnormal Lung Sounds in Limited Data Setting», en *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Mexico: IEEE, nov. 2021, pp. 527-530. doi: 10.1109/EMBC46164.2021.9630091.

[34] L. Cohen, *Time-frequency analysis*. en Prentice Hall signal processing series. Englewood Cliffs, N.J: Prentice Hall PTR, 1995.

[35] R. L. Allen y D. W. Mills, *Signal Analysis: Time, Frequency, Scale, and Structure*, 1.ª ed. Wiley, 2003. doi: 10.1002/047166037X.

[36] B. P. Lathi, *Linear Systems and Signals*, 2.ª ed. Oxford University Press, 2010.

[37] S. Haykin y B. Van Veen, *Signals and Systems*, 2.ª ed. Wiley, 2002.

[38] R. K. Hobbie y B. J. Roth, *Intermediate Physics for Medicine and Biology*. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-12682-1.

[39] N. E. Huang *et al.*, «The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis», *Proc. R. Soc. Lond. A*, vol. 454, n.º 1971, pp. 903-995, mar. 1998, doi: 10.1098/rspa.1998.0193.

[40] S. Kadambe y G. F. Boudreaux-Bartels, «A comparison of the existence of "cross terms" in the Wigner distribution and the squared magnitude of the wavelet transform and the short-time Fourier transform», *IEEE Trans. Signal Process.*, vol. 40, n.º 10, pp. 2498-2517, oct. 1992, doi: 10.1109/78.157292.

[41] Ö. Şayli, «Wheeze detection in the respiratory sounds using Hilbert-Huang transform.», *22nd Signal Processing and Communications Applications Conference (SIU)*, 2014, doi: 10.1109/SIU.2014.6830699.

[42] M. Lozano, J. A. Fiz, y R. Jane, «Automatic Differentiation of Normal and Continuous Adventitious Respiratory Sounds Using Ensemble Empirical Mode Decomposition and Instantaneous Frequency», *IEEE J. Biomed. Health Inform.*, vol. 20, n.º 2, pp. 486-497, mar. 2016, doi: 10.1109/JBHI.2015.2396636.

[43] M. Lozano, J. A. Fiz, y R. Jané, «Performance evaluation of the Hilbert–Huang transform for respiratory sound analysis and its application to continuous adventitious sound characterization», *Signal Processing*, vol. 120, pp. 99-116, mar. 2016, doi: 10.1016/j.sigpro.2015.09.005.

[44] M. A. Colominas, G. Schlotthauer, y M. E. Torres, «Improved complete ensemble EMD: A suitable tool for biomedical signal processing», *Biomedical Signal Processing and Control*, vol. 14, pp. 19-29, nov. 2014, doi: 10.1016/j.bspc.2014.06.009.

[45] M. E. Torres, M. A. Colominas, G. Schlotthauer, y P. Flandrin, «A complete ensemble empirical mode decomposition with adaptive noise», en *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic: IEEE, may 2011, pp. 4144-4147. doi: 10.1109/ICASSP.2011.5947265.

[46] P. Dehkordi, A. Garde, B. Molavi, J. M. Ansermino, y G. A. Dumont, «Extracting Instantaneous Respiratory Rate From Multiple Photoplethysmogram Respiratory-Induced Variations», *Front. Physiol.*, vol. 9, p. 948, jul. 2018, doi: 10.3389/fphys.2018.00948.

[47] H.-T. Wu, Y.-H. Chan, Y.-T. Lin, y Y.-H. Yeh, «Using synchrosqueezing transform to discover breathing dynamics from ECG signals», *Applied and Computational Harmonic Analysis*, vol. 36, n.º 2, pp. 354-359, mar. 2014, doi: 10.1016/j.acha.2013.07.003.

[48] S. Meignen, T. Oberlin, y D.-H. Pham, «Synchrosqueezing transforms: From low- to high-frequency modulations and perspectives», *Comptes Rendus Physique*, vol. 20, n.º 5, pp. 449-460, jul. 2019, doi: 10.1016/j.crhy.2019.07.001.

[49] W. Ertel, *Introduction to Artificial Intelligence*. en Undergraduate Topics in Computer Science. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-58487-4.

[50] F. Chollet, *Deep learning with Python*. Shelter Island, New York: Manning Publications Co, 2018.

[51] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, y G. Zhang, «Transfer learning using computational intelligence: A survey», *Knowledge-Based Systems*, vol. 80, pp. 14-23, may 2015, doi: 10.1016/j.knosys.2015.01.010.

[52] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, y K. Keutzer, «SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size», 2017.

[53] «Get Started with Transfer Learning», Help Center, MATLAB, The MathWorks, Inc. [En línea]. Disponible en: https://la.mathworks.com/help/deeplearning/gs/get-started-with-transfer-learning.html?lang=en

[54] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[55] J. Delpiano, «Fast mutual information of two images or signals.», MATLAB Central File Exchange. [En línea]. Disponible en: https://www.mathworks.com/matlabcentral/fileexchange/13289-fast-mutual-information-of-two-images-or-signals

[56] A. P. Córdova-Martínez, S. Charleston-Villalobos, T. Aljama-Corrales, y B. A. Reyes, «Assessment of an automated wheezing respiratory cycle identification based on CNN and classical and high-resolution time-frequency representations.».