ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Resumen

El propósito de esta investigación es analizar un método para obtener parámetros acústicos a partir de señales de voz grabadas en una sala, evitando así la necesidad de medir las respuestas al impulso del entorno. Para lograr esto, se propone un modelo basado en redes neuronales convolucionales, inspirado en las técnicas más avanzadas en el campo. A diferencia de los enfoques anteriores, la singularidad de esta tesis radica en la capacidad de obtener múltiples parámetros simultáneamente.

Para entrenar esta red, se utiliza la información de la envolvente temporal de amplitud (TAE) de una señal de voz reverberada. A partir de esta señal, se extraen cuatro descriptores que corresponden a los parámetros buscados: tiempo de reverberación (T_{30}), claridad (C_{50} y C_{80}) y definición (D_{50}). Dado el limitado número de datos disponibles, se genera una base de datos de grabaciones de voz reverberada convolucionando señales de voz anecoicas con una variedad de respuestas al impulso reales, generadas artificialmente y aumentadas a partir de impulsos reales.

Finalmente, para evaluar la viabilidad del modelo propuesto, se comparan los descriptores obtenidos en una sala utilizando métodos convencionales con los resultados del modelo. Se lleva a cabo un análisis de correlación entre las diferencias esperadas y los valores obtenidos utilizando el conjunto de datos de prueba. Excluyendo la banda de 125 Hz, se obtienen valores de correlación promedio superiores al 93 % para el descriptor T_{30}, demostrando su alta precisión. Además, los descriptores restantes, C_{80}, D_{50} y C_{50}, presentaron valores de correlación por encima del 87 % y del 67 % respectivamente. Estos resultados indican que la red neuronal logra estimar eficazmente los cuatro parámetros acústicos en diversas bandas.

Palabras Clave: respuestas al impulso, tiempo de reverberación, parámetros acústicos de una sala, envolvente temporal de amplitud.

Abstract

The purpose of this research is to analyze a method to obtain acoustic parameters from speech signals recorded in a room, thus avoiding the need to measure the impulse responses of the environment. To achieve this, a model based on convolutional neural networks is proposed, inspired by the most advanced techniques in the field. Unlike previous approaches, the uniqueness of this thesis lies in the ability to obtain multiple parameters simultaneously.

To train this network, the information of the temporal amplitude envelope (TAE) of a reverberated speech signal is used. From this signal, four descriptors are extracted corresponding to the parameters sought: reverberation time (T_{30}), clarity (C_{50} and C_{80}) and definition (D_{50}). Given the limited amount of data available, a database of reverberated speech recordings is generated by convolving anechoic speech signals with a variety of real impulse responses, artificially generated impulse responses and augmented from real impulses.

Finally, to assess the feasibility of the proposed model, the descriptors obtained in a room using conventional methods are compared with the results of the model. A correlation analysis is carried out between the expected differences and the values obtained using the test data set. Excluding the 125 Hz band, average correlation values above 93 % are obtained for the T_{30} descriptor, demonstrating its high accuracy. In addition, the remaining descriptors, C_{80}, D_{50} and C_{50}, presented correlation values above 87 % and 67 %, respectively. These results indicate that the neural network manages to effectively estimate the four acoustic parameters in various bands.

Keywords: impulse responses, reverberation time, room acoustic parameters, temporal amplitude envelope.

1. INTRODUCCIÓN

1.1. FUNDAMENTACIÓN

La medición de parámetros acústicos desempeña un papel fundamental en el estudio del comportamiento del campo sonoro en recintos. Existen diversos parámetros destinados a representar características acústicas relevantes para el diseño de estos espacios [1]. En general, el objetivo final al diseñar un recinto está vinculado a valores específicos de dichos parámetros, los cuales deben respetarse para lograr una óptima respuesta sonora en relación al uso previsto [2][3][4].

Existen técnicas ampliamente conocidas y estandarizadas para obtener los parámetros objetivos, como el índice de transmisión de voz (STI), el tiempo de reverberación (RT), la claridad (C_{80}), entre otros [5][6]. Estos descriptores, así como la mayoría de otros no mencionados, se calculan a partir de la respuesta al impulso de una sala. Esto se debe a que, según la teoría de señales y sistemas, un sistema lineal e invariante en el tiempo (LTI) queda completamente definido por su respuesta al impulso. En general, podemos considerar que un recinto con una fuente emisora fija y un punto receptor se comporta como un sistema de estas características. Por lo tanto, al tener la respuesta al impulso de un recinto (RIR), podemos comprender el comportamiento de cualquier estímulo al atravesar dicho sistema [7].

Existen diversos métodos, algunos más sofisticados que otros, para calcular la respuesta al impulso de una sala. El procedimiento de obtención implica generar una señal de corta duración y alta amplitud que pueda excitar el recinto en todas las frecuencias. Entre los métodos más conocidos se encuentran el disparo en blanco (disparo de un arma de fuego que contiene cartuchos con pólvora y casquillos, pero sin proyectil), la explosión de globos, los aplausos, el choque con maderas, el barrido frecuencial, entre otros [8]. Sin embargo, todos estos métodos se ven afectados por los ruidos presentes en la sala, ya sean generados por personas o provenientes de maquinaria, tráfico u otras fuentes. Esta problemática es especialmente relevante en situaciones en las que no es posible desalojar la sala, como en una estación de subterráneo. Además, muchos de estos métodos tienen limitaciones en cuanto al espectro de frecuencias que pueden excitar, como el caso de la explosión de globos, donde el diámetro del globo determina la frecuencia más baja en la que puede generar energía

[9]. Por otra parte, la mayoría de estos métodos no son replicables, lo que puede generar resultados diferentes al repetir el mismo procedimiento.

En los últimos años, se ha desarrollado la metodología del barrido frecuencial para abordar problemas relacionados con el piso de ruido, el rango frecuencial excitado y la repetitibilidad de las mediciones [10]. Esta metodología ha logrado superar estas limitaciones. Sin embargo, es importante destacar que su implementación requiere de equipamiento costoso y de difícil transporte, lo que puede representar una desventaja en ciertos contextos. Por esta razón, en la práctica actual del estudio de recintos, aún se siguen utilizando los métodos mencionados anteriormente debido a su practicidad de implementación, a pesar de sus limitaciones.

Por lo tanto, en el marco de esta investigación, se propone un método novedoso para la medición de parámetros acústicos. Este enfoque consiste en estimar dichos parámetros a partir de una señal de voz grabada que contiene la reverberación propia del recinto en cuestión. Esta técnica se conoce en la literatura especializada como estimación ciega de parámetros acústicos, ya que, a diferencia de los métodos convencionales, no requiere el uso de equipos para medir directamente la respuesta al impulso de la sala. En cambio, se basa en la modelación de la reverberación utilizando algoritmos basados en redes neuronales y el procesamiento de un audio de voz reverberado [11].

Para lograr este objetivo, la investigación se enmarca en el campo de la ingeniería de audio, centrándose en el estudio y comprensión de las limitaciones de los modelos existentes. Se plantea mejorar estos modelos y llevar a cabo un análisis detallado de los resultados obtenidos. Mediante este enfoque, se busca avanzar en el desarrollo de soluciones más efectivas y precisas para la estimación de parámetros acústicos a partir de señales de voz reverberadas.

1.2. OBJETIVOS

1.2.1. Objetivo General

El objetivo general de esta investigación es la implementación de un algoritmo basado en redes neuronales convolucionales que permita estimar los parámetros acústicos de un recinto utilizando una grabación de señal de voz reverberada como entrada. El propósito es desarrollar un enfoque novedoso y efectivo para la estimación de estos parámetros, utili-

zando técnicas de aprendizaje profundo y aprovechando las capacidades de las redes neuronales convolucionales en el procesamiento de señales acústicas. El algoritmo propuesto busca superar las limitaciones de los enfoques convencionales y ofrecer una solución más precisa y eficiente en la estimación de los parámetros acústicos de una sala.

1.2.2. Objetivos Específicos

Entre los objetivos específicos de la investigación se incluyen:

- Revisar las diferentes técnicas utilizadas para obtener parámetros acústicos cuando no se dispone de la respuesta al impulso de la sala.
- Diseñar e implementar un modelo utilizando redes neuronales en el lenguaje de programación Python, que permita la obtención ciega de parámetros acústicos a partir de señales de voz reverberadas.
- Generar una base de datos de grabaciones de voz con reverberación utilizando tanto respuestas impulsivas reales como generadas artificialmente.
- Aplicar técnicas de aumentación de datos para mejorar la diversidad y representatividad de las respuestas al impulso reales utilizadas en el modelo.
- Optimizar el sistema propuesto y comparar los resultados obtenidos con los parámetros acústicos calculados mediante métodos convencionales.
- Realizar mediciones empíricas de una sala para contrastar y validar los resultados obtenidos por el modelo propuesto.

1.3. ESTRUCTURA DE LA INVESTIGACIÓN

En el capítulo 2 se realiza un estudio exhaustivo sobre el estado del arte en el campo de la estimación ciega de parámetros acústicos. Se revisa la literatura existente y se analizan las técnicas y enfoques utilizados en investigaciones previas, destacando las fortalezas y limitaciones de los métodos existentes.

En el capítulo 3 se presenta el marco teórico necesario para comprender el trabajo. Se describe en detalle la respuesta al impulso de una sala (RIR) y las técnicas de síntesis de respuestas al impulso. También se abordan las técnicas de aumentación de datos aplicadas

a las respuestas al impulso, así como los parámetros acústicos a estimar, como T_{30}, C_{50}, C_{80} y D_{50}. Además, se introduce el concepto de envolvente temporal de amplitud (TAE) y se explora la aplicación de redes neuronales convolucionales y algoritmos de aprendizaje en el procesamiento de señales acústicas.

En el capítulo 4 se especifica la metodología seguida a lo largo del trabajo y se brinda toda la información necesaria para replicar los experimentos realizados. Se detalla la recopilación de datos, la preparación de la base de datos, el diseño del modelo y la configuración de los experimentos. También se describe el procedimiento de entrenamiento y evaluación del modelo propuesto, así como las métricas utilizadas para la evaluación de los resultados.

En el capítulo 5 se presentan los resultados de los experimentos y se realiza un análisis crítico de los mismos. Se exponen los hallazgos obtenidos y se discuten en comparación con los métodos convencionales. Se busca interpretar los resultados y resaltar su relevancia en el campo de la estimación ciega de parámetros acústicos.

En el capítulo 6 se exponen las conclusiones generales del trabajo, se reflexiona sobre las limitaciones y se plantean posibles mejoras. También se destaca la importancia y las implicancias de los resultados obtenidos.

En el capítulo 7 se proponen líneas futuras de investigación relacionadas con el presente estudio, se sugieren posibles extensiones, aplicaciones y enfoques prometedores que podrían ser explorados en el futuro.

2. ESTADO DEL ARTE

2.1. MODELOS DE ESTIMACIÓN CIEGA DE PARÁMETROS ACÚSTICOS

En la actualidad, existen diversos métodos ampliamente conocidos para la obtención de parámetros acústicos que permiten caracterizar una sala. Estos están detallados en el marco de diversas normas reconocidas en el campo. Entre los parámetros más comunes y utilizados se encuentran: EDT, T_{10}, T_{20}, T_{30}, C_{50}, C_{80}, D_{50}, Ts [5], STI [6],entre otros .

Todos los parámetros acústicos mencionados anteriormente, así como muchos otros, se calculan a partir de la respuesta al impulso de la sala (RIR) [5]. Sin embargo, existen diferentes técnicas disponibles para medir esta respuesta, y cada una tiene sus ventajas y limitaciones.

Las técnicas de bajo costo de implementación suelen presentar limitaciones en términos de su alcance en frecuencia, ruido de fondo y repetibilidad. Entre estas se encuentran métodos como el disparo en blanco, la explosión de globos, los aplausos o el choque con maderas, entre otros. Aunque son prácticos y económicos, pueden generar resultados inconsistentes debido a la presencia de ruido y a su limitado espectro de frecuencias. Esto puede ser especialmente problemático en situaciones donde se requiere una medición precisa de la respuesta al impulso, como en salas de conciertos o estudios de grabación.

Por otro lado, existen técnicas más sofisticadas que abordan las limitaciones mencionadas. Estas suelen requerir equipamiento especializado y costoso para generar estímulos de manera precisa y controlada. Entre estas se destacan el barrido frecuencial, la utilización de fuentes de ruido de banda ancha y la utilización de micrófonos y altavoces de alta calidad. Estas técnicas permiten obtener mediciones más precisas y confiables de la respuesta al impulso, abarcando un amplio rango de frecuencias y minimizando la influencia del ruido de fondo.

En resumen, existe un compromiso entre el costo de implementación y las capacidades de las técnicas disponibles para medir la respuesta al impulso de una sala. Las de bajo costo son prácticas pero pueden ser limitadas en términos de precisión y confiabilidad, mientras que las más sofisticadas ofrecen resultados más precisos pero requieren un equipamiento costoso y especializado.

En respuesta a las limitaciones en la obtención de la respuesta al impulso de una sala

(RIR), algunos investigadores han propuesto métodos de estimación ciega de parámetros acústicos. Uno de los primeros avances en este campo fue presentado por Kendrick et al. en 2007 [12], donde se propuso un método para estimar el tiempo de reverberación de una sala a partir de una grabación de audio realizada en el lugar. Este enfoque se denominó estimación ciega debido a que no se requería la medición directa de la RIR.

En la norma ISO 3382 [5], el tiempo de reverberación se calcula a partir de la pendiente de la curva de decaimiento de la RIR, utilizando cuadrados mínimos para aproximar la disminución en decibeles en rangos específicos (-5 a -15 para T_{10}, -5 a -25 para T_{20} y -5 a -35 para T_{30}). Los autores descubrieron una correlación entre el decaimiento de la RIR y el final de una palabra en un audio de voz o en un acorde tocado en una sala. Basándose en esta observación, grabaron audios reverberados capturados en la sala y seleccionaron fragmentos con curvas de decaimiento similares a las de la RIR. Luego, aplicaron un estimador de máxima verosimilitud (MLE) para aproximar estas curvas a una pendiente ideal y calcular el tiempo de reverberación de banda completa de la sala con buen nivel de precisión.

Este enfoque novedoso permitió estimar el tiempo de reverberación sin la necesidad de medir directamente la RIR, utilizando grabaciones de audio y técnicas de procesamiento de señales.

El método anteriormente descrito tiene algunas limitaciones, como la incapacidad para estimar el tiempo de reverberación por bandas de frecuencia, la influencia del ruido de fondo y la falta de capacidad para determinar otros parámetros acústicos. Con estas consideraciones en mente, los mismos autores presentaron un modelo adicional al año siguiente que aborda estas limitaciones [13].

En este nuevo método, aprovechando el auge de las redes neuronales artificiales, se entrena una red utilizando la envolvente de grabaciones de voz realizadas en una sala y filtradas en una banda de frecuencia específica. La salida de esta red es el valor que se intenta aproximar, correspondiente al descriptor acústico de la sala en la banda de frecuencia previamente filtrada. De esta manera, el sistema es capaz de calcular los parámetros T_{20}, EDT, C_{80} y Ts por bandas de frecuencia, superando las limitaciones del método anterior y proporcionando una estimación más precisa y completa de los parámetros acústicos.

Hasta ese momento, la estimación ciega de parámetros acústicos era factible, pero el método presentaba limitaciones. Este modelo se encontraba altamente influenciado por

el ruido de fondo presente en las salas, lo que afectaba la precisión de las estimaciones. Además, las redes neuronales utilizadas en el proceso no eran eficientes, ya que estimaban un único parámetro a la vez, lo que requería entrenar una red para cada parámetro deseado.

En ese contexto, se presentó el *Acoustic Characterisation of Environments Challenge* (ACE Challenge) en 2015, organizado por el *IEEE Audio and Acoustic Signal Processing Technical Committee* [14]. Este desafío tenía como objetivo evaluar algoritmos de vanguardia para la estimación ciega del tiempo de reverberación (RT) y la relación directo-reverberado (DRR) de una sala a partir de grabaciones de voz, y promover la investigación en este campo en constante evolución.

Para el desafío, se proporcionó a los participantes una base de datos de respuestas al impulso grabadas en cinco salas y grabaciones de voz anecoicas que podrían ser reverberadas [15]. Los participantes debían desarrollar algoritmos capaces de estimar el RT y el DRR utilizando únicamente las grabaciones de voz reverberadas sin conocer las respuestas al impulso reales.

Este desafío motivó la presentación de diversos modelos innovadores para la estimación del RT y DRR. Entre ellos, se destacaron los enfoques propuestos por Parada et al. [16], que utilizaron redes neuronales recurrentes para estimar estos parámetros en segmentos pequeños de audio, aprovechando la capacidad de memoria de estas redes. Otro enfoque interesante fue el presentado por Prego et al. [17], que calcula los parámetros utilizando el decaimiento de la señal de voz filtrada en múltiples bandas de frecuencia y promediando los valores de RT y DRR obtenidos. Por último, se encuentra el método propuesto por Loellman et al. [18], que emplea un estimador de máxima verosimilitud (MLE) para estimar los valores de RT y DRR. Estos últimos dos métodos pueden considerarse como una extensión del enfoque presentado por Kendrick et al. [12]. Estos modelos representaron avances significativos en la estimación ciega de parámetros acústicos, mejorando la precisión y la aplicabilidad de estos métodos en diversas aplicaciones relacionadas con el análisis y caracterización acústica de espacios.

No obstante, la falta de una base de datos extensa y diversa de respuestas al impulso (RIR) se identificó como una limitación en los modelos existentes, ya que no podían abordar eficazmente audios con baja relación señal-ruido ni grabaciones de salas no incluidas en el conjunto de entrenamiento. Para superar este desafío, Bryan propuso un método de

aumentación de respuestas al impulso, que permite generar una mayor cantidad de RIRs y así modelar una variedad más amplia de salas con menos esfuerzo [19]. La utilización de esta nueva base de datos mejoró significativamente los modelos existentes propuestos por otros autores, tanto para audios con alto nivel de ruido como para grabaciones tomadas en salas no contempladas en el conjunto de entrenamiento. Esta estrategia de aumentación de datos demostró ser una solución efectiva para abordar las limitaciones anteriores y mejorar la generalización de los modelos de estimación ciega de parámetros acústicos.

Hasta este punto, la mayoría de los modelos existentes se centraban en estimar un solo parámetro a la vez, como el tiempo de reverberación (RT) o la relación directo-reverberado (DRR). Sin embargo, en 2021, Duangpummet et al. propusieron un método innovador [11] que permite calcular varios parámetros acústicos simultáneamente a partir de estimar únicamente el RT. En este enfoque, se utiliza una red neuronal que se entrena con la envolvente de una señal de audio grabada en la sala. La red estima el RT por bandas de frecuencia y, a partir de este valor, sintetiza una respuesta al impulso (RIR) suponiendo que puede ser representada como una exponencial decreciente con un decaimiento determinado por el tiempo de reverberación. Con la obtención de esta RIR sintética, el modelo puede calcular todos los descriptores acústicos utilizando los métodos convencionales establecidos en la norma ISO 3382. Este enfoque permite calcular múltiples parámetros acústicos de manera eficiente y precisa a partir de una única señal de audio.

Si bien este último modelo es revolucionario por tener la capacidad de calcular varios descriptores a la vez, asume una hipótesis que no es cierta en todos los casos. Al reconstruir la respuesta al impulso asumiendo una pendiente ideal, se pueden obtener resultados imprecisos, ya que existen múltiples respuestas al impulso que podrían tener el mismo tiempo de reverberación pero características distintas en otros parámetros. Esta suposición simplificada puede llevar a estimaciones incorrectas cuando la correlación entre el tiempo de reverberación y otros descriptores no es fuerte.

En la práctica, la pendiente ideal de una respuesta al impulso se obtiene únicamente en salas con geometría paralelepipédica, donde su decaimiento puede ser modelado utilizando una señal de ruido modulada por una exponencial decreciente. Sin embargo, cuando los recintos presentan otras geometrías, sus RIRs exhiben comportamientos más complejos debido a que las reflexiones de orden superior pueden disminuir a velocidades diferentes

a las de orden inferior. En tales casos, se requiere el uso de múltiples exponenciales para modelar adecuadamente la curva de decaimiento.

Para lograr resultados más precisos y confiables, sería necesario que la respuesta al impulso reconstruida proporcione información más completa y realista sobre las características de la sala. Esto implicaría capturar detalles adicionales más allá del tiempo de reverberación, de modo que la envolvente de la respuesta sintética se asemeje más a la que se obtendría en una medición real dentro de esa sala.

3. MARCO TEÓRICO

3.1. RESPUESTA AL IMPULSO DE UNA SALA: RIR

Según la teoría de señales y sistemas, cualquier sistema lineal e invariante en el tiempo (LTI) puede describirse mediante su respuesta al impulso. Esta propiedad es de gran utilidad en numerosas disciplinas de la ingeniería, incluida la acústica.

Consideremos un conjunto que consiste en una sala equipada con un micrófono y una fuente de sonido. Para caracterizar el recinto y comprender sus propiedades acústicas, se busca obtener la respuesta al impulso *h(t)*. Esto permite calcular parámetros que revelan información relevante sobre la sala. Para obtener la respuesta al impulso, se requiere excitar el sistema con un impulso delta de Dirac, que es un estímulo de duración infinitamente corta. Al realizar esta excitación en diferentes ubicaciones de la sala, se obtienen distintos valores de respuesta al impulso.

Es importante destacar que este enfoque asume que el sistema se comporta como un sistema lineal e invariante en el tiempo, aunque en la práctica esto no siempre es estrictamente válido. Sin embargo, esta suposición no afecta significativamente el cálculo de los descriptores acústicos. Aunque existen algunas variaciones y limitaciones en determinadas situaciones, el modelo de respuesta al impulso sigue siendo una herramienta eficaz para comprender las características acústicas de una sala.

En este caso, como se utiliza un micrófono para captar la respuesta al impulso del sistema, los métodos de excitación deben ser acústicos. Algunos de los métodos comunes son:

- Disparo en blanco
- Explosión de globos
- Aplausos
- Choque con maderas
- Barrido frecuencial logarítmico (LSS, del inglés *logarithmic sine sweep*)

Entre estos métodos, el LSS ha ganado popularidad debido a sus ventajas. A diferencia de los demás, este ofrece un mayor control y repetibilidad de la medición, ya que permite

ajustar la duración, el rango de frecuencias y la amplitud del estímulo. Otra ventaja es que se pueden realizar múltiples mediciones y promediarlas para mejorar la relación señal-ruido de la respuesta al impulso. Además, el LSS permite evaluar la distorsión introducida por el sistema de medición utilizado [10].

A modo ilustrativo, se muestra en la Figura 1 un ejemplo de la respuesta al impulso de una sala generada de forma sintética. La generación artificial de respuestas al impulso se abordará más adelante en este mismo capítulo.

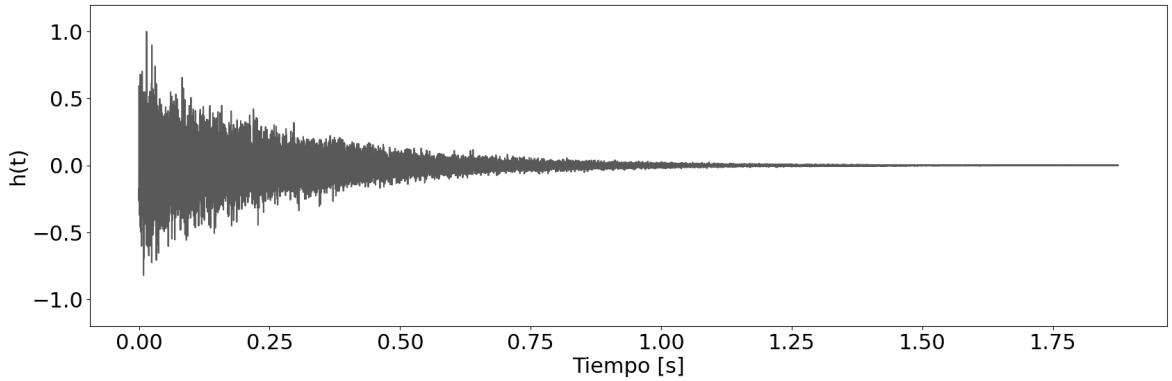

Figura 1. Ejemplo de respuesta al impulso de una sala.

Al considerar la sala como un sistema LTI y conocer su respuesta al impulso, es posible obtener la señal que sería captada por un micrófono ubicado en un punto específico del recinto al ser estimulado desde su interior. Matemáticamente, esto se representa mediante la convolución entre la respuesta al impulso de la sala y la señal emitida por la fuente, como se muestra en la ecuación 1.

$$y(t) = h(t) * s(t) \tag{1}$$

Donde *h(t)* es la respuesta al impulso de la sala, *s(t)* es el estímulo o señal de entrada, e *y(t)* es la respuesta del sistema, a la cual se denomina señal reverberada.

En otras palabras, si se dispone de un estímulo anecoico (libre de reverberación), es posible convolucionarlo con la respuesta al impulso de la sala y obtener como resultado una señal que incorpora todas las características acústicas del recinto, como si hubiera sido grabada desde el interior del mismo.

3.2. DESCRIPTORES ACÚSTICOS DE UNA SALA

En esta sección, se definen los parámetros acústicos utilizados en la presente investigación, los cuales se basan en las definiciones establecidas por la norma ISO 3382 [5] para los descriptores de tiempo de reverberación (T_{30}), claridad (C_{50}, C_{80}) y definición (D_{50}).

3.2.1. Tiempo De Reverberación: EDT, T_{10}, T_{20}, T_{30} Y T_{60}

El tiempo de reverberación es uno de los descriptores más relevantes y ampliamente utilizados para caracterizar un recinto acústico. Se define como el tiempo necesario para que la densidad de energía sonora, promediada en todo el espacio de la sala, disminuya en 60 dB una vez que la fuente de sonido ha dejado de emitir. Este parámetro se conoce comúnmente como T_{60}.

En la mayoría de las situaciones reales, las condiciones de ruido en una sala no permiten que la fuente de sonido emita una señal con un rango dinámico mayor a 60 dB desde el nivel máximo hasta el nivel de ruido de fondo necesario para calcular el descriptor. Por lo tanto, surge la necesidad de realizar el cálculo en rangos más limitados y luego proyectar los resultados a la disminución deseada de 60 dB.

Por lo tanto, para calcular el descriptor T_{60} a partir de una respuesta al impulso conocida, se utilizan diferentes rangos de la curva de decaimiento suavizada. Estos rangos se definen en relación al máximo de la señal y se corresponden con los descriptores EDT, T_{10}, T_{20} y T_{30}, según las siguientes condiciones:

- EDT: Se toma el rango de la curva desde el máximo de la señal hasta 10 dB por debajo del mismo.
- T_{10}: Se considera el rango de la curva desde 5 dB hasta 15 dB por debajo del máximo.
- T_{20}: Se utiliza el rango de la curva desde 5 dB hasta 25 dB por debajo del máximo.
- T_{30}: Se emplea el rango de la curva desde 5 dB hasta 35 dB por debajo del máximo.

Existen diferentes métodos para suavizar la señal de decaimiento de una RIR, entre ellos se encuentran el filtrado de media móvil [20], la transformada de Hilbert [21] y el método de integración de Schroeder [22]. En esta investigación se opta por utilizar el método

de integración de Schroeder debido a su eficacia y amplia aceptación en la comunidad científica.

El método de integración de Schroeder se basa en el cuadrado de la RIR, escalado en decibelios. La ecuación 2 muestra la forma de calcular la versión suavizada $h_s(t)$ de la RIR $h(t)$.

$$h_s(t) = \frac{\int_t^\infty h^2(t)dt}{\int_0^\infty h^2(t)dt} \tag{2}$$

Es fundamental tener en cuenta que este método no contempla la presencia de ruido en la respuesta al impulso. Por lo tanto, se asume un límite de integración hasta infinito, lo cual no refleja la realidad donde el ruido es inherente y variable en las mediciones acústicas. Para abordar esta limitación, se requiere la implementación de un criterio de recorte de la curva de decaimiento.

En este estudio, se adopta el método de Lundeby [23] para determinar el nivel de ruido presente en la señal y establecer un límite superior de integración en la fórmula de Schroeder. Este método proporciona una estimación confiable del piso de ruido, lo que permite ajustar el rango de integración de acuerdo con las características específicas de cada medición.

Al combinar el método de integración de Schroeder con el enfoque de Lundeby, se logra obtener una envolvente suavizada que refleja con mayor precisión el decaimiento de la RIR, permitiendo así un cálculo más confiable de los descriptores acústicos.

Una vez obtenida la curva de decaimiento suavizada, es posible calcular los descriptores acústicos de tiempo de reverberación, como EDT, T_{10}, T_{20} y T_{30}, utilizando la pendiente de la recta aproximada por cuadrados mínimos dentro del rango correspondiente mencionado anteriormente.

Por lo tanto, considerando que m representa la pendiente de la recta de aproximación por cuadrados mínimos, los descriptores de tiempo de reverberación se determinan utilizando la ecuación 3.

$$T_x = \frac{60}{m} \tag{3}$$

Aquí, T_x representa los descriptores EDT, T_{10}, T_{20} o T_{30} según el rango utilizado para

estimar la pendiente de la recta.

A modo ilustrativo, se muestra en la Figura 2 la curva de decaimiento de una RIR, junto con las pendientes aproximadas mediante cuadrados mínimos correspondientes a cada descriptor.

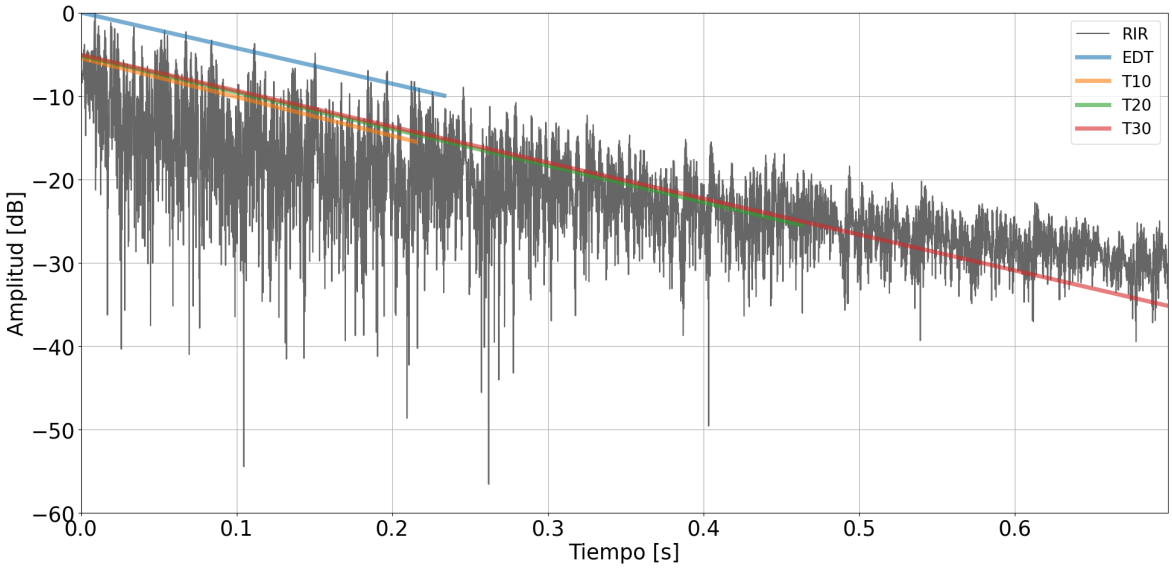

Figura 2. Decaimiento de una respuesta al impulso y sus aproximaciones por cuadrados mínimos para los descriptores EDT, T10, T20 y T30.

3.2.2. Claridad: C_{80} Y C_{50}

La claridad es un descriptor que se utiliza para comprender la relación entre la energía directa y la energía tardía en una sala, lo que proporciona información sobre el campo sonoro antes y después de la llegada de las reflexiones al micrófono. Específicamente, el descriptor C_{80} se utiliza para determinar la transparencia de las salas de música.

Dada una respuesta al impulso $h(t)$, el descriptor C_{80} (expresado en dB) se calcula dividiendo la energía en los primeros 80 ms de la señal por la energía en el resto de la señal, como se muestra en la ecuación 4.

$$C_{80} = 10Log_{10}\left[\frac{\int_0^{80ms} h^2(t)dt}{\int_{80ms}^{\infty} h^2(t)dt}\right] \tag{4}$$

Además de C_{80}, existe otro descriptor llamado C_{50} que se utiliza para determinar la transparencia del habla en una sala. Su cálculo es similar solo que cuenta con límites de integración diferentes, que van de 0 a 50 ms en lugar de 80 ms.

3.2.3. Definición: D_{50}

La definición es otro descriptor utilizado para evaluar la respuesta de una sala al comparar la energía en la señal temprana con la energía en la señal tardía.

Específicamente, el D_{50} se utiliza para medir la inteligibilidad del habla en salas de conferencias o aulas, y se expresa como un porcentaje.

Dada una respuesta al impulso $h(t)$, la definición se puede calcular utilizando la ecuación 5.

$$D_{50} = 100\frac{\int_0^{50ms} h^2(t)dt}{\int_0^{\infty} h^2(t)dt} \tag{5}$$

3.3. RESPUESTAS AL IMPULSO SINTÉTICAS

Para modelar una respuesta al impulso (RIR) estocástica, es posible utilizar el método de Schroeder [24]. Este método supone que la señal puede ser representada por un ruido blanco gaussiano cuya amplitud está modulada por una exponencial decreciente.

Por tanto, la respuesta al impulso por banda de frecuencia se puede modelar utilizando la ecuación 6.

$$h(t) = ae^{\frac{-6,9t}{T_{60}}} c_h(t) \tag{6}$$

Donde a es la amplitud inicial, T_{60} es el tiempo de reverberación de la sala y $c_h(t)$ es un ruido blanco gaussiano.

Luego, conociendo las respuestas al impulso por bandas de frecuencia, la RIR original puede ser representada como la suma de todas las bandas. Esto se observa en la ecuación 7.

$$h(t) = \sum_{k=1}^{n} e^{\frac{-6,9t}{T_{60,k}}} c_{h,k}(t) \tag{7}$$

Donde $T_{60,k}$ es el tiempo de reverberación en la k-ésima banda (en un total de n bandas) y $c_{h,k}(t)$ es un ruido blanco gaussiano limitado en frecuencia.

En la sección 4.2.2, se ampliarán los conceptos abordados en la ecuación 6, presentando un modelo de mayor generalidad diseñado para facilitar la implementación de técnicas

de aumentación del tiempo de reverberación.

Como se mencionó anteriormente, este modelo solo es capaz de representar RIRs de salas paralelepipédicas y sin obstáculos internos, lo cual no es representativo de los casos de recintos reales. Por tanto, se necesita definir un modelo realista para la caída de la presión sonora dentro de una habitación.

Sabiendo que las reflexiones de orden superior pueden decaer a velocidades diferentes que las de orden inferior, debido a que los campos de sonido en las habitaciones no son completamente difusos en la mayoría de los casos, es posible plantear un modelo de síntesis que utilice el decaimiento de más de una exponencial para contemplar este fenómeno [12].

Siguiendo la lógica de las ecuaciones anteriores, la respuesta al impulso se puede modelar como una señal de envolvente ($e(t)$) multiplicada por un ruido blanco gausiano ($c_h(t)$), como se observa en la ecuación 8.

$$h(t) = e(t)c_h(t) \tag{8}$$

Ahora, la envolvente se representa como una suma de envolventes, como se observa en la ecuación 9.

$$e(t) = \sum_{k=1}^{M} \alpha_k \beta_k^t \tag{9}$$

Donde β_k representa las tasas de decaimiento, α_k son los pesos de los factores y M es la cantidad de decaimientos considerados.

En la mayoría de los casos, todas las RIRs se pueden modelar utilizando únicamente dos curvas de decaimiento [12]. Por tanto, la ecuación de la envolvente puede simplificarse como se observa en la ecuación 10.

$$e(t) = \alpha_k \beta_1^t + (1-\alpha)\beta_2^t \tag{10}$$

3.4. RELACIÓN DIRECTO-REVERBERADO: DRR

Dada la respuesta al impulso de una sala (RIR), el descriptor DRR se define, para una posición específica del recinto, como la relación entre el nivel de presión sonora de un sonido directo proveniente de una fuente direccional y el nivel de presión sonora reverberante que

incide simultáneamente en el mismo punto [25]. Por consiguiente, es dependiente de la distancia entre el punto emisor y receptor y del tiempo de reverberación del recinto.

Este parámetro se puede calcular matemáticamente utilizando la ecuación 11.

$$DRR = 10Log_{10}\left(\frac{\sum_{n=o}^{n_d} h^2[n]}{\sum_{n=n_d+1}^{\infty} h^2[n]}\right) \tag{11}$$

Donde $h[n]$ representa la respuesta al impulso en tiempo discreto y n_d corresponde a las muestras correspondientes a la señal directa. Por tanto, todas las muestras que continúan luego de n_d corresponden al campo reverberado causado por las reflexiones de la onda sonora en el recinto.

Teniendo en cuenta las muestras n_d, la respuesta al impulso $h[n]$ se divide en dos componentes distintas: la parte que corresponde al sonido directo, denominada $h_e[n]$, y el campo reverberante, denominado $h_l[n]$.

3.5. ENVOLVENTE TEMPORAL DE AMPLITUD: TAE

La envolvente temporal de amplitud (TAE) es una representación simplificada de una señal de voz captada en un recinto, que se enfoca únicamente en la información de la envolvente de la señal en lugar de su contenido completo. Basándonos en la lógica del cálculo de descriptores acústicos mediante una respuesta al impulso (RIR), sabemos que la envolvente de la señal es lo más relevante y no el contenido específico de la misma. Por lo tanto, resulta útil generar una versión simplificada de la señal de voz que contenga únicamente datos de la envolvente temporal.

Al estudiar la TAE de la señal de voz en lugar de la señal sin procesar, se obtienen varias ventajas. En primer lugar, se logra independizarse del contenido del audio, ya que lo único importante en el análisis son las partes de decaimiento de la señal, de las cuales se pueden obtener los parámetros acústicos deseados. En segundo lugar, la información se comprime drásticamente, lo que resulta en una disminución en la complejidad del entrenamiento de redes neuronales u otros algoritmos utilizados en el análisis de señales acústicas. Esto facilita el procesamiento de la señal y agiliza los cálculos necesarios para extraer los descriptores acústicos relevantes.

Para obtener la TAE se deben realizar los siguientes pasos:

1. **Obtener una señal reverberada.**

2. **Filtrar la señal con un filtro pasa banda:** siendo esta alguna de las bandas centrales comprendidas entre 125 Hz y 8 kHz

3. **Obtener la magnitud de la transformada de Hilbert de la señal:** esta devuelve la envolvente compleja de una señal, de la cual luego se extrae su amplitud.

4. **Filtrar la señal con un filtro pasa bajos con frecuencia de corte en 20 Hz:** se aplica un filtro que elimina las frecuencias altas de la envolvente de la señal, ya que en el contexto de la TAE no es necesario conservar el contenido detallado de la señal de voz, sino únicamente su envolvente temporal.

5. **Remuestrear la señal a una frecuencia de muestreo de 40 Hz:** dado que la señal ya ha sido filtrada y no contiene información relevante por encima de 20 Hz, se puede reducir su frecuencia de muestreo a 40 Hz, lo cual disminuye la cantidad de muestras y facilita el procesamiento y entrenamiento de modelos.

6. **Normalizar la señal obtenida.**

Los pasos descriptos anteriormente se pueden observar de forma gráfica en la Figura 3.

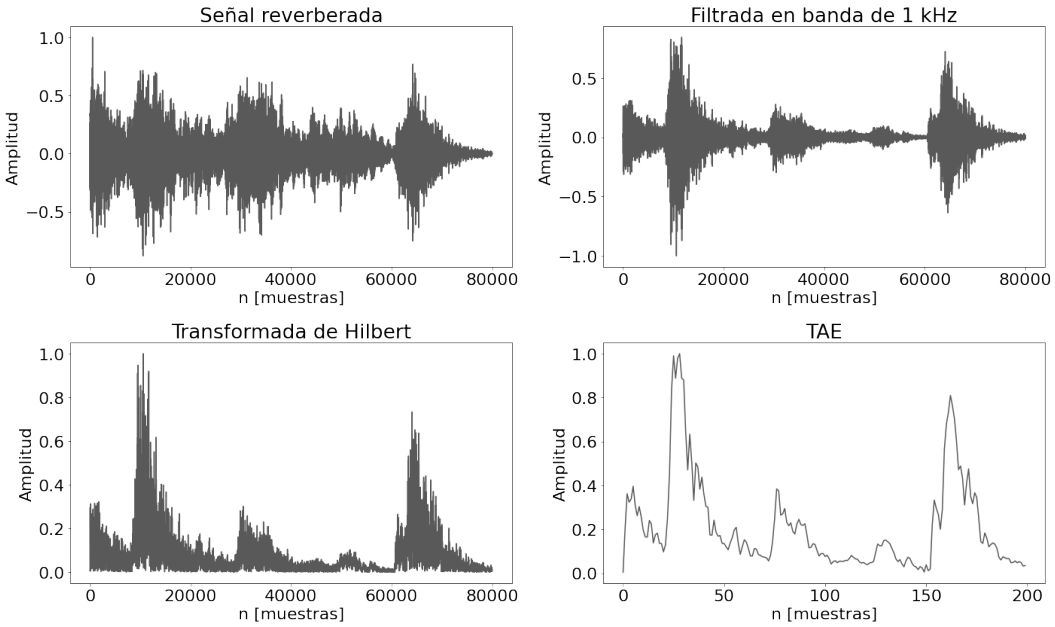

Figura 3. Pasos para la obtención de la TAE.

3.6. REDES NEURONALES ARTIFICIALES

3.6.1. La Neurona Artificial O Perceptrón

Dentro del campo de investigación del aprendizaje automático, también conocido como machine learning, las redes neuronales destacan como uno de los algoritmos fundamentales [26].

Para comprender el funcionamiento de este algoritmo, es necesario estudiar su unidad básica, la neurona. Existen diversos modelos de neuronas, pero el más reconocido y ampliamente utilizado es el perceptrón [27].

El perceptrón, como una forma específica de neurona artificial, representa la unidad mínima en una red neuronal. Su objetivo principal es realizar cálculos que permitan detectar características y patrones en los datos de entrada.

El esquema básico de un perceptrón se puede observar en la Figura 4.

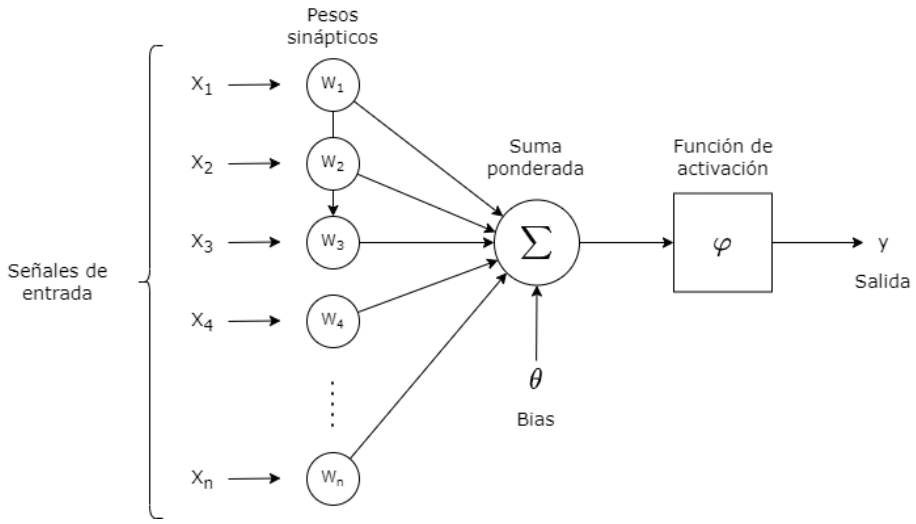

Figura 4. Esquema de una neurona artificial.

La figura anterior ilustra los componentes principales de una neurona artificial, que son los siguientes:

- **Entradas:** Representan los datos que serán procesados por la neurona.
- **Pesos sinápticos:** Son parámetros asociados a cada entrada que se ajustan durante el entrenamiento de la red neuronal. Estos pesos determinan la importancia relativa de cada entrada en el cálculo.

- **Umbral o Bias:** Es una entrada adicional a la neurona y su valor también se ajusta durante el entrenamiento. El umbral influye en la activación de la neurona.

- **Suma ponderada:** En esta etapa, se multiplican los valores de las entradas por sus pesos correspondientes, y luego se suma el resultado junto con el valor del umbral. Esta operación se puede representar mediante la ecuación 12.

$$\sum_{i=1}^{n} X_i W_i + \theta \tag{12}$$

- **Función de activación:** Después de la suma ponderada, se aplica una función de activación a la salida de la neurona. Esta función determina si la neurona se activa o no. La ecuación 13 muestra el cálculo general de la salida de la neurona, donde φ representa la función de activación. Algunos ejemplos comunes de funciones de activación se ilustran en la Figura 5.

$$y = \varphi\left(\sum_{i=1}^{n} X_i W_i + \theta\right) \tag{13}$$

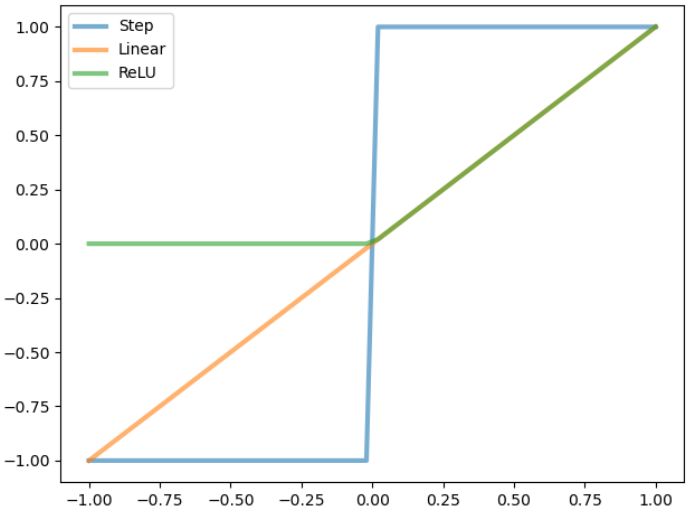

Figura 5. Funciones de activación.

3.6.2. Perceptrón Multicapa

Una vez comprendido el funcionamiento de una neurona artificial, el siguiente paso es combinar varias de ellas para formar lo que se conoce como una red neuronal.

Una red neuronal es un conjunto de neuronas interconectadas que se organizan en capas. Cada capa consta de un número determinado de neuronas, y las salidas de las neuronas de una capa suelen ser las entradas de las neuronas de la capa siguiente. Estas redes se construyen siguiendo una arquitectura específica, que implica definir la cantidad de capas, el número de neuronas por capa, el tipo de capas, entre otros parámetros.

Uno de los modelos de redes neuronales más estudiados es el perceptrón multicapa [28]. Esta red neuronal consta de una capa de entrada que se conecta a una o varias capas ocultas, y una capa de salida. Cada capa puede tener un número diferente de neuronas y está completamente conectada a la capa adyacente. Cada neurona tiene asociados pesos y un umbral. Si la salida de cualquier neurona individual supera un umbral especificado, la neurona se activa y envía datos a la siguiente capa de la red. De lo contrario, no se transmiten datos a la siguiente capa.

La Figura 6 muestra un ejemplo de este tipo de redes.

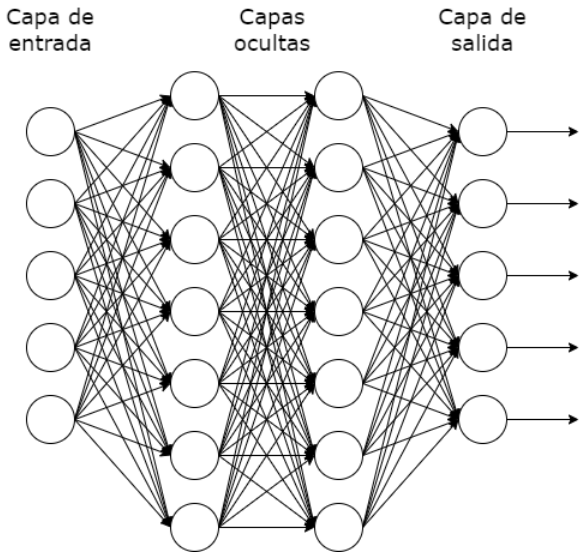

Figura 6. Perceptrón multicapa.

Una de las ventajas más significativas de este algoritmo es su capacidad para representar cualquier función matemática si la red tiene un número suficiente de neuronas. Esta propiedad se conoce como el teorema de aproximación universal de las redes neuronales [29].

En particular, cuando estas redes neuronales tienen dos o más capas ocultas, se les suele llamar redes de aprendizaje profundo [30]. Este tipo de redes profundas han demos-

trado ser especialmente eficaces en tareas de aprendizaje automático y han impulsado avances significativos en áreas como la visión por computadora, el procesamiento de lenguaje natural y muchas otras aplicaciones.

3.6.3. Entrenamiento De Una Red Neuronal

Los algoritmos de aprendizaje automático se basan en el procesamiento de datos para entrenar modelos. En el caso del aprendizaje supervisado, se utilizan datos de entrenamiento en forma de pares (x, y), donde y representa el valor objetivo esperado para una determinada entrada x.

El objetivo de una red neuronal es encontrar una configuración óptima de los parámetros entrenables que permita que una entrada x genere la salida deseada y. A este proceso se le conoce como entrenamiento. A medida que la red se somete a un entrenamiento adecuado, la salida generada por la red se acerca cada vez más al valor objetivo, lo que indica que la red está aprendiendo.

Durante el proceso de aprendizaje, es natural que las salidas de la red difieran inicialmente de los valores objetivo, ya que los parámetros de la red se inicializan de manera aleatoria. Por lo tanto, es esencial contar con una métrica que cuantifique esta diferencia. La función de costo se encarga de medir esta discrepancia. Esta función compara las salidas de la red con las salidas esperadas y calcula un error utilizando una función matemática específica.

El objetivo del entrenamiento es minimizar la función de costo, es decir, reducir el error entre las salidas de la red y los valores objetivo. Para lograrlo, se utilizan algoritmos de optimización que ajustan iterativamente los parámetros de la red, actualizando su configuración en función del error calculado. Este proceso se repite varias veces, abarcando todo el conjunto de datos de entrenamiento, hasta que la red neuronal pueda generar salidas cercanas a los valores objetivo para la mayoría de las entradas.

El resultado de la función de costo se utiliza como una señal de retroalimentación para ajustar los parámetros entrenables de la red neuronal y minimizar el error. Esta tarea es llevada a cabo por otra función conocida como función de optimización. La función de optimización utiliza el algoritmo de retropropagación del error para calcular el gradiente de la función de costo con respecto a los parámetros entrenables de la red. Utilizando este

gradiente y el valor de la tasa de aprendizaje definida en la función de optimización, se determina cómo modificar los parámetros entrenables para reducir el error de salida.

Las herramientas utilizadas para controlar el proceso de aprendizaje de una red neuronal se dividen en dos grupos: parámetros e hiperparámetros. Los parámetros son valores que se obtienen a través del entrenamiento, mientras que los hiperparámetros se utilizan para configurar el proceso de instanciación del modelo y describir su configuración.

Algunos de los hiperparámetros que deben definirse para el entrenamiento de una red incluyen el tamaño de los lotes y las épocas. Los lotes se refieren a la segmentación del conjunto de datos de entrenamiento. El entrenamiento se realiza en iteraciones, donde en cada iteración se procesa un lote de datos. Una vez que todos los lotes han sido procesados, se completa una época.

Además de los hiperparámetros y los parámetros entrenables, la forma y el tipo de datos que se presentan a la red tienen un impacto significativo en su rendimiento. El objetivo es lograr que la red neuronal pueda generalizar y estimar de manera precisa datos que no ha visto previamente. Para lograr esto, la base de datos se divide en los siguientes conjuntos:

- **Conjunto de entrenamiento:** este conjunto es el más grande y se utiliza durante la etapa de entrenamiento de la red neuronal.
- **Conjunto de validación:** este conjunto se utiliza para evaluar el rendimiento del sistema durante la etapa de entrenamiento.
- **Conjunto de prueba:** este conjunto se utiliza para evaluar el rendimiento de la red neuronal al realizar estimaciones. Contiene datos que nunca han sido vistos por la red.

Cada uno de estos conjuntos cumple un papel importante en el proceso de entrenamiento y evaluación de la red neuronal, permitiendo evaluar su capacidad de generalización y su rendimiento en datos no vistos previamente.

Es importante mencionar dos fenómenos no deseados que pueden ocurrir en las redes neuronales cuando no están correctamente implementadas: el sobreajuste y el subajuste. El sobreajuste se produce cuando el modelo muestra un rendimiento excelente en los datos de entrenamiento, pero un rendimiento deficiente en los datos de validación y evaluación.

Esto indica que el modelo no ha logrado generalizar el problema, sino que simplemente ha memorizado los datos de entrenamiento. Por otro lado, el subajuste se refiere a un rendimiento deficiente tanto en los datos de entrenamiento como en los de evaluación.

El sobreajuste ocurre cuando la complejidad del modelo es demasiado alta en comparación con la tarea que se está tratando de resolver, o cuando el tamaño del conjunto de datos de entrenamiento es insuficiente. Por el contrario, el subajuste se produce cuando el modelo es demasiado simple en comparación con la complejidad de la tarea que intenta resolver.

Para abordar un conjunto de datos de entrenamiento insuficiente, una técnica efectiva es la de aumentación de datos. Esta técnica consiste en generar nuevos datos a partir de los datos existentes. Al agregar casos de prueba adicionales durante el entrenamiento, se mejora la capacidad de generalización del modelo. Si la base de datos no es capaz de representar todos los casos relevantes del problema, es posible manipular la información existente para ampliarla y generar nuevos datos basados en los originales.

Durante el entrenamiento de una red neuronal, existen varias técnicas que pueden utilizarse para mejorar los resultados. En esta investigación, se emplean tres técnicas específicas: la normalización por lotes (*batch normalization*), el *dropout* y el *max pooling*.

La normalización por lotes se utiliza para normalizar los valores de salida de una capa de neuronas. Consiste en calcular la media y la desviación estándar de la salida de una capa en un lote de datos y luego normalizar la salida restando la media y dividiendo por la desviación estándar. Esta técnica ayuda a estabilizar el proceso de entrenamiento al mantener la distribución de los datos en un rango adecuado.

El *dropout* se utiliza para mitigar el sobreajuste en las redes neuronales. Durante el entrenamiento, se aplica una operación de apagado de manera aleatoria a un cierto porcentaje de las neuronas en una capa. Esto implica que las neuronas seleccionadas no contribuirán a la propagación hacia adelante ni a la propagación hacia atrás en esa iteración en particular. Al desactivar aleatoriamente algunas neuronas, la red se ve obligada a aprender características más robustas y generalizables, ya que no puede depender en exceso de un pequeño conjunto de neuronas.

Por último, el *max pooling* es una técnica comúnmente utilizada en el procesamiento digital de imágenes. Su objetivo es reducir la dimensión espacial de una capa de caracterís-

ticas, disminuyendo así la cantidad de parámetros y cálculos realizados por la red neuronal. El *max pooling* divide la capa de entrada en regiones y selecciona el valor máximo de cada región como salida. Esto permite conservar las características más destacadas y relevantes de la capa de entrada, mientras reduce su tamaño.

Estas técnicas son herramientas efectivas para mejorar el rendimiento y la capacidad de generalización de una red neuronal durante el entrenamiento. Al aplicarlas de manera adecuada, se puede obtener un modelo más robusto y preciso en la tarea que se desea resolver.

3.6.4. Redes Neuronales Convolucionales (CNN)

Las redes neuronales convolucionales son un tipo de red neuronal profunda en el que las neuronas corresponden a campos receptivos, similar al funcionamiento de la corteza visual. Este algoritmo utiliza filtros convolucionales para extraer características significativas de las imágenes, como bordes, texturas, formas, entre otros. Estos filtros se aplican repetidamente a la imagen, reduciendo su tamaño a medida que se profundiza en la red. Luego, la información resultante se alimenta a capas completamente conectadas para que la red pueda realizar una tarea específica, como la clasificación de la imagen [31].

La capa convolucional es el componente fundamental de las redes neuronales convolucionales. Presenta varias características distintivas que la hacen adecuada para el procesamiento de datos en forma de cuadrícula, como imágenes. A continuación, se describen las características principales de la capa convolucional:

- **Campo receptivo limitado**: En lugar de estar conectadas a todas las entradas, las neuronas de la capa convolucional están conectadas solo a una región local de la entrada. Esta región se denomina campo receptivo y permite a la capa convolucional capturar patrones locales en los datos de entrada. A medida que se profundiza en la red, el campo receptivo se expande para abarcar regiones más amplias de la entrada.

- **Parámetros compartidos**: En una capa convolucional, los pesos sinápticos de las neuronas se comparten entre diferentes ubicaciones del campo receptivo. Esto significa que un mismo conjunto de pesos se utiliza para calcular la activación de múltiples ubicaciones en la entrada. Esta propiedad de compartir parámetros reduce drásticamente la cantidad de parámetros entrenables en la red y permite que las redes

convolucionales sean más eficientes en términos de memoria y cálculo. Además, al compartir parámetros, las redes convolucionales adquieren la capacidad de detectar patrones en diferentes ubicaciones de la entrada, lo que les otorga invariancia a la traslación.

- **Convolución**: La operación central en una capa convolucional es la convolución. La convolución se realiza aplicando un filtro de convolución a la entrada. El filtro, también conocido como kernel, es una pequeña matriz de pesos. Durante la convolución, el filtro se desliza sobre la entrada realizando productos internos entre sus valores y una sección correspondiente de la entrada. Estos productos internos se suman para obtener un único valor de salida. El proceso de convolución se repite en diferentes ubicaciones de la entrada para generar un mapa de características, que representa la presencia de ciertos patrones en la señal. Cada filtro aprende a detectar un patrón específico, como bordes, texturas o formas, en la entrada. La Figura 7 ilustra este proceso.

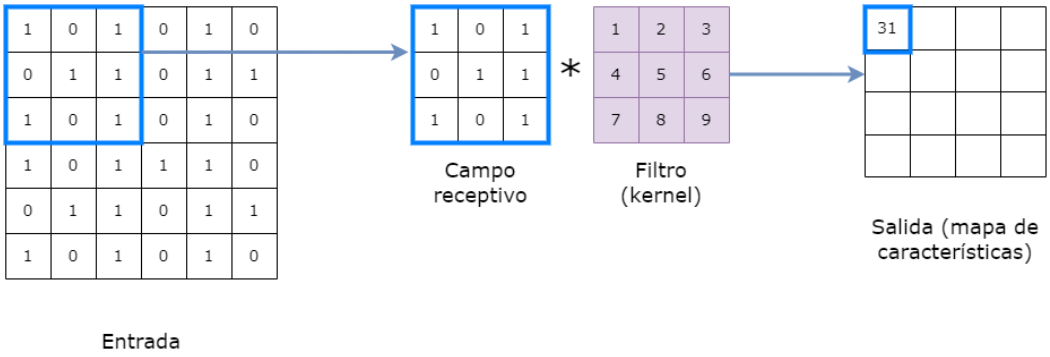

Figura 7. Esquema de un filtro de convolución.

En las redes neuronales convolucionales, se incorporan hiperparámetros adicionales que no se encuentran presentes en las redes de perceptrón multicapa. Estos hiperparámetros permiten ajustar el comportamiento y la arquitectura de la red para adaptarse a las características específicas del problema. A continuación, se describen estos hiperparámetros:

- **Tamaño del filtro**: Este hiperparámetro define el tamaño espacial del campo receptivo utilizado en la convolución. Los filtros convolucionales suelen tener una forma

cuadrada, como 3x3 o 5x5, y determinan el patrón específico que se busca detectar en la entrada.

- **Tamaño del salto (stride)**: Este hiperparámetro indica la cantidad de desplazamiento horizontal y vertical que se aplica al filtro convolucional durante la convolución. El tamaño del salto determina la distancia entre los campos receptivos adyacentes en la capa convolucional. Un tamaño de salto mayor reduce el tamaño de la salida y puede ayudar a reducir la cantidad de cálculos necesarios, pero también puede perder detalles finos en la entrada.

- **Relleno de ceros (zero-padding)**: El relleno de ceros es una técnica utilizada para mantener las dimensiones de entrada y salida de una capa convolucional. Consiste en agregar una cantidad de ceros alrededor de la entrada antes de aplicar la convolución. El relleno de ceros ayuda a preservar el tamaño de la imagen y a mantener la información en los bordes, ya que sin él, la convolución tendería a reducir gradualmente el tamaño de la imagen en las capas posteriores.

- **Cantidad de filtros aplicados**: Este hiperparámetro determina la cantidad de filtros convolucionales que se aplican a la entrada. Cada filtro convolucional aprende a detectar un conjunto de características específicas en la entrada. La cantidad de filtros aplicados es igual al número de mapas de características generados, y cada mapa de características representa la presencia de un patrón diferente en la señal.

En resumen, las redes neuronales convolucionales utilizan estos hiperparámetros adicionales para controlar el tamaño del filtro, el tamaño del salto, el relleno de ceros y la cantidad de filtros aplicados. Estos hiperparámetros permiten a las redes convolucionales generalizar conceptos visuales complejos con una cantidad menor de parámetros en comparación con una red completamente conectada. La distribución estratégica de estos hiperparámetros a lo largo de la arquitectura de la red es clave para lograr un procesamiento eficiente y efectivo de los datos de entrada.

4. METODOLOGÍA

4.1. GENERACIÓN DE BASE DE DATOS

Para lograr que una red neuronal aprenda tareas complejas, es fundamental disponer de una base de datos extensa que sea capaz de representar todas las características del fenómeno que se desea modelar. Además, es importante garantizar que todos los elementos de la base de datos compartan los mismos tipos de codificación, como la misma profundidad de bits, frecuencia de muestreo y formato de audio.

En el contexto de esta investigación, la base de datos se compone de envolventes temporales de amplitud (TAE) de audios de voz reverberados, junto con los valores de los descriptores acústicos T_{30}, C_{50}, C_{80} y D_{50} de la sala en la que se generó la reverberación del audio. Se tomó la decisión de almacenar las TAEs en lugar de los audios reverberados para reducir significativamente el tamaño de la base de datos, ya que cada TAE consta únicamente del 0.25 % de la cantidad de muestras del audio original que se extrae.

La generación de la base de datos se llevó a cabo utilizando audios de voz anecoicos grabados en la cámara anecoica de la Universidad Tecnológica de Delft [32]. Estos audios se sometieron a procesos de reverberación mediante la convolución con una serie de respuestas al impulso de diferentes salas, lo que generó variaciones en la base de datos. Para cada audio reverberado y su correspondiente respuesta al impulso, se calcularon tanto los valores de TAE como los descriptores acústicos.

Una vez generada la base de datos, se procedió a dividirla en tres grupos: conjunto de entrenamiento, validación y prueba. El grupo de prueba comprende el 20 % del total de la base de datos, mientras que el conjunto de entrenamiento representa el 70 % y el de validación el 10 % restante.

Con el fin de garantizar un entrenamiento adecuado y una evaluación posterior precisa de la red, se aseguró que las TAE generadas para la base de datos provinieran de archivos diferentes, es decir, se evitó la repetición de archivos entre los conjuntos de entrenamiento y prueba. Para lograr esto, se realizó una separación de las respuestas al impulso de manera que el 80 % de cada grupo (reales, aumentadas y sintéticas) se utilizara para el entrenamiento y validación, y el 20 % restante se reservara para el conjunto de pruebas. De esta manera, se garantizó que las respuestas al impulso utilizadas en el grupo de prueba no fueran vistas

previamente por la red durante su entrenamiento. Además, se logró un conjunto de pruebas homogéneo en el cual se compararon todos los escenarios. Asimismo, la base de datos de voz se segmentó de manera que no se repitieran los mismos archivos entre los diferentes grupos y se mantuviera cierta homogeneidad en cada uno de ellos. Los detalles específicos de esta segmentación se abordan en la sección 4.5.

4.2. AUMENTACIÓN DE RESPUESTAS AL IMPULSO

En el contexto de esta investigación, es necesario contar con una amplia variedad de respuestas al impulso que reflejen diferentes características acústicas, como los tiempos de reverberación y las pendientes de decaimiento de diferentes salas. Sin embargo, obtener estas respuestas al impulso a través de mediciones in situ resulta complicado y costoso.

Una estrategia para abordar este desafío es utilizar técnicas de aumentación de datos. Estas técnicas implican aplicar transformaciones a las señales de respuesta al impulso existentes con el objetivo de modificar sus parámetros acústicos, como el tiempo de reverberación (T_{30}) y la relación directo-reverberado (DRR). De esta manera, es posible aumentar significativamente la cantidad de salas modeladas en el conjunto de datos.

Es importante destacar que en esta investigación, las respuestas al impulso aumentadas se obtienen a partir del procesamiento de RIRs reales. Para lograrlo, se aplican dos procesos específicos: la modificación de la amplitud en la parte inicial de la respuesta al impulso para controlar la relación directo-reverberado, y la alteración de la envolvente de decaimiento para controlar el tiempo de reverberación [19].

4.2.1. Aumentación De La Relación Directo-Reverberado

Para abordar la modificación del descriptor DRR, es importante tener en cuenta que este descriptor considera dos componentes de la respuesta al impulso: el sonido directo y el campo reverberado, denotados como $h_e(t)$ y $h_l(t)$, respectivamente (según se explicó en la sección 3.4).

Para alterar este descriptor, se aplica una ganancia específica a la parte correspondiente al sonido directo (que, convencionalmente, abarca los primeros 2.5 ms de la respuesta al impulso). Esta ganancia está determinada por un factor α y se ajusta para obtener el valor deseado del descriptor DRR. Como resultado, se obtiene una nueva señal llamada $\tilde{h}_e(t)$. Sin

embargo, este proceso por sí solo puede generar discontinuidades en la señal. Para evitarlo, se aplican ventanas complementarias. Esto implica calcular una señal directa ventaneada y un residuo ventaneado, como se muestra en la ecuación 14.

$$\tilde{h}_e(t) = \alpha w_d(t) h_e(t) + [1 - w_d(t)] h_e(t) \tag{14}$$

Donde, $w_d(t)$ representa una venta Hann de 5 ms de longitud.

Al combinar las ecuaciones 14 y 11, se plantea un sistema de ecuaciones en el que se establece un valor deseado para el descriptor DRR y se resuelve para obtener el valor correspondiente de α.

En la Figura 8, se ilustra el proceso de aumentación del descriptor DRR. Se muestra la representación de una parte del campo directo $h_e(t)$, las ventanas aplicadas, el efecto del factor α y la señal resultante $\tilde{h}_e(t)$.

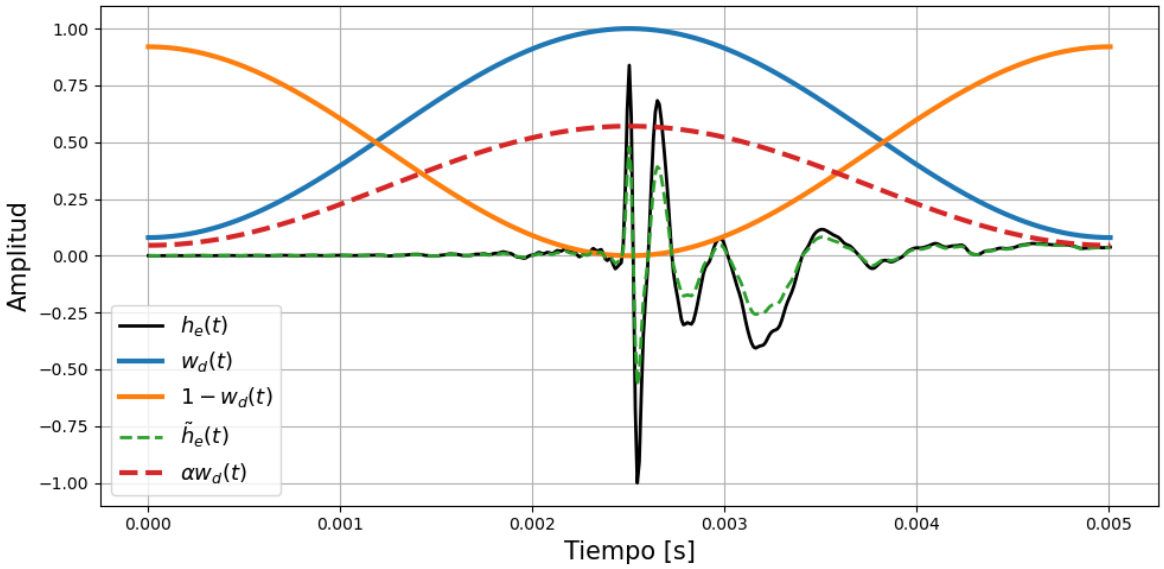

Figura 8. Proceso de aumentación del DRR.

Posteriormente, la parte modificada del campo directo se concatena con el resto de la respuesta al impulso, generando así la nueva respuesta aumentada.

Es importante destacar que, para tiempos de reverberación muy cortos, no siempre es posible obtener relaciones directo-reverberado muy bajas, ya que la energía tardía de la respuesta al impulso es inicialmente muy baja. Para abordar esta situación, se definen límites para el valor de α aplicado.

4.2.2. Aumentación Del Tiempo De Reverberación

A diferencia del caso del DRR, la aumentación del descriptor T_{30} se enfoca en la parte tardía de la respuesta al impulso.

En la sección 3.3 se estudió que la respuesta al impulso puede ser modelada como un ruido blanco Gaussiano con una caída de nivel dominada por una función exponencial decreciente. Esta caída exponencial depende de la frecuencia y se asocia a un nivel de ruido específico.

Considerando esto, se utiliza el modelo de síntesis de respuestas al impulso descrito en la ecuación 15 para la aumentación.

$$h_m(t) = A_m e^{\frac{-(t-t_0)}{\tau_m}} c_h(t) u(t-t_0) + \sigma_m c_h(t) \tag{15}$$

En esta ecuación, A_m representa la amplitud inicial, τ_m es la tasa de decaimiento, σ_m es el nivel de ruido de fondo, $c_h(t)$ es un ruido Gaussiano con media cero y desviación estándar uno, t_0 es el tiempo de inicio de la parte tardía de la respuesta al impulso, m es el índice que indica la sub-banda en la que se está trabajando, y $u(t)$ es una función escalón unitario.

En este modelo, el tiempo de reverberación se relaciona con el parámetro τ mediante la ecuación 16.

$$T_{60} = \ln(1000)\tau T_s \tag{16}$$

Aquí, T_s representa el período de muestreo.

Se aplican métodos de optimización no lineales al modelo de RIR definido en la ecuación 15 para estimar los parámetros $\hat{A}_m$, $\hat{\tau}_m$ y $\hat{\sigma}m$ que mejor se ajusten a la envolvente de decaimiento de la respuesta al impulso. Luego, utilizando estos parámetros y calculando la tasa de decaimiento para la banda deseada $\tau_{m,d}$ (donde d indica la banda que se está evaluando) basada en el tiempo de reverberación buscado, se modifica la parte tardía de la RIR original multiplicándola por una envolvente exponencial creciente o decreciente, según corresponda. Esta modificación se muestra en la ecuación 17.

$$h'_m(t) = h_m(t) e^{-(t-t_0)\frac{\hat{\tau}_m - \tau_{m,d}}{\hat{\tau}_m \tau_{m,d}}} \tag{17}$$

Aquí, $h'_m(t)$ representa la nueva parte tardía generada de la respuesta al impulso.

El proceso de alteración del tiempo de reverberación consiste en modificar la pendiente de decaimiento de la RIR hasta alcanzar la deseada para cada banda de frecuencia. Una vez obtenidas las respuestas para cada sub-banda, se suman para obtener la respuesta final.

El proceso descrito anteriormente funciona únicamente cuando se desea sintetizar tiempos de reverberación más cortos que los de la señal original. En estos casos, se multiplica la respuesta al impulso por una exponencial decreciente. Sin embargo, cuando se busca generar tiempos de reverberación más largos, es necesario multiplicar por una exponencial creciente, lo que conlleva una amplificación de la señal. Al hacerlo, también se amplifica el nivel de ruido presente en la señal, lo que resulta en pendientes de decaimiento inestables que no se corresponden con la respuesta original.

Para evitar este problema, es necesario estimar el nivel de ruido de la respuesta al impulso. Se utiliza el método de Lundeby [23] para llevar a cabo esta tarea. Una vez estimado el nivel de ruido, se obtiene la respuesta final realizando una transición suave (cross-fade) entre la parte tardía generada y una cola reverberante sintética. Esta cola se obtiene multiplicando ruido gaussiano por una envolvente exponencial decreciente, utilizando los parámetros calculados previamente.

Resumiendo, los pasos a seguir para llevar a cabo la aumentación del tiempo de reverberación son los siguientes:

- Normalizar la respuesta al impulso.
- Filtrar la respuesta en bandas de octava.
- Estimar el nivel de ruido presente en la respuesta al impulso.
- Estimar la envolvente de decaimiento.
- Sintetizar una señal utilizando la envolvente estimada y un nivel de ruido nulo.
- Realizar un cross-fade entre la señal sintetizada y la respuesta original en el punto de inicio del nivel de ruido.
- Multiplicar la señal por una exponencial creciente o decreciente según corresponda.

- Sumar las respuestas de las diferentes bandas para obtener la respuesta sintetizada en todo el espectro.
- Combinar la parte tardía aumentada con la parte directa de la respuesta al impulso original.

4.3. CURVA DE DECAIMIENTO DE UNA SEÑAL DE VOZ

Retomando la idea de considerar una sala con un micrófono y una fuente de sonido como un sistema, podemos entender que cualquier señal emitida en su interior se verá afectada por la respuesta al impulso del recinto, como se muestra en la ecuación 1. En otras palabras, la señal captada en el interior de la sala contendrá información acerca de las características de la misma.

Basándose en esta premisa, los investigadores Kendrick et al. [12] observaron en su investigación que la curva de decaimiento al final de una palabra presenta cierta similitud con la respuesta al impulso de una sala. Además, esta curva se ve influenciada por la reverberación presente en la sala, lo que resulta en un tiempo de decaimiento prolongado debido al tiempo de reverberación del recinto.

Este fenómeno es de gran importancia, ya que al grabar una oración dentro de la sala, los investigadores pudieron determinar el tiempo de reverberación de la sala mediante el cálculo de la media de los tiempos de reverberación extraídos de las diferentes curvas de decaimiento presentes en la señal de audio captada. Esta fue la primera técnica utilizada para estimar el tiempo de reverberación de una sala de manera indirecta.

A modo ilustrativo, en la Figura 9 se compara la curva de decaimiento de una señal sin reverberación con la misma señal afectada por la reverberación de la sala.

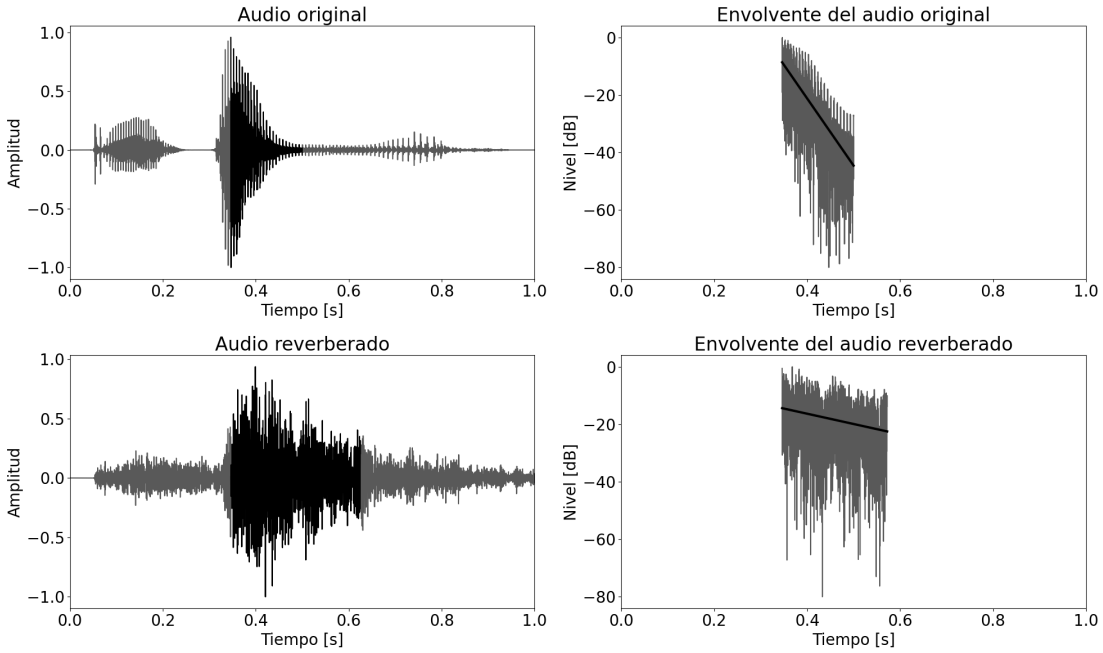

Figura 9. Diferencias en la curva de decaimiento para un audio limpio contra uno reverberado.

4.4. OBTENCIÓN DE RESPUESTAS AL IMPULSO

En el marco de esta investigación, se utilizaron dos tipos de respuestas al impulso: reales y sintéticas. Además, se generó un tercer grupo de respuestas mediante el proceso de aumentación aplicado a las RIR reales.

En todos los casos, las señales se muestrearon a una frecuencia de 16 kHz. Este aspecto es crucial, ya que se buscó homogeneizar la codificación de la base de datos con el fin de evitar problemas durante el entrenamiento de la red neuronal.

4.4.1. Respuestas Al Impulso Sintéticas

Para la generación de respuestas al impulso sintéticas, se empleó el modelo de Schroeder, representado por la ecuación 7. Este modelo permite sintetizar RIRs de salas con forma de paralelepípedo, dado un tiempo de reverberación específico. La señal se obtiene multiplicando un ruido blanco gaussiano de media cero y desviación estándar unitaria por una exponencial decreciente.

Específicamente, se generaron un total de 2900 respuestas al impulso de manera equitativa con respecto al tiempo de reverberación. Los valores de RT abarcaron el rango de 0.2 s a 3 s, con incrementos de 0.1 s. Se seleccionó este rango para mantener consistencia con el trabajo previo [11]. En cuanto a la duración de las señales, se decidió que fueran 0.5 s más

largas que el tiempo de reverberación utilizado en su síntesis.

A modo de ilustración, la Figura 10 muestra tres ejemplos de respuestas al impulso generadas utilizando el modelo de Schroeder con diferentes tiempos de reverberación.

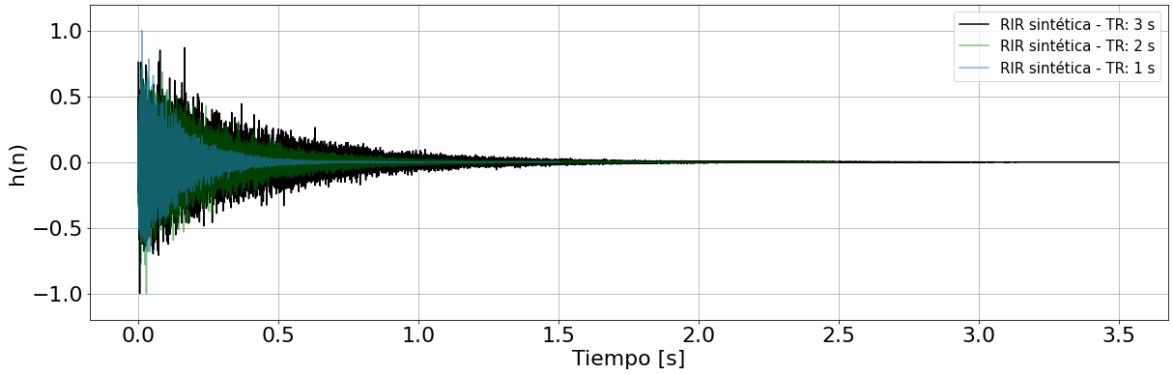

Figura 10. Respuestas al impulso con distintos tiempos de reverberación generadas de forma sintética.

4.4.2. Respuestas Al Impulso Reales

Las respuestas al impulso reales utilizadas en este estudio se obtuvieron de la base de datos C4DM [33]. Esta base de datos fue seleccionada debido a que está acompañada de un documento detallado que describe la metodología empleada para su adquisición, así como los planos de los recintos medidos. Además, la base de datos C4DM es de fácil acceso y está disponible en su página web oficial.

Las RIRs se obtuvieron utilizando la técnica de barrido frecuencial (LSS) [10]. Para la generación de las señales, se utilizó un altavoz Genelec 8250A como fuente y un micrófono omnidireccional modelo DPA 4006 para la captura. El altavoz empleado es un transductor de dos vías, con un driver de 8’ para frecuencias bajas y medias, y uno de 1’ para frecuencias altas.

En concreto, se realizaron mediciones en tres recintos para generar este conjunto de datos. El primero corresponde a una sala multipropósito con aproximadamente 800 asientos (*Great Hall*). El segundo es una biblioteca de estilo victoriano (*Octagon*). Por último, se utilizó un salón de clases de una universidad (*Classroom*).

Tanto en la sala multipropósito como en la biblioteca se calcularon un total de 169 respuestas al impulso, mientras que en el salón de clases se midieron 130. En todos los casos, se utilizó una única posición de la fuente de sonido y se variaron las posiciones de los micrófonos. Estas posiciones de los micrófonos se distribuyeron en una grilla equiespaciada

dentro de los recintos, lo que permitió obtener una cobertura uniforme del espacio.

En la Tabla 1 se presentan los tiempos de reverberación correspondientes a cada recinto.

Tabla 1. Tiempos de reverberación por bandas de frecuencia para cada sala.

	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz	8000 Hz
Great Hall [s]	$2{,}19 \pm 1{,}71$	$2{,}17 \pm 0{,}16$	$2{,}44 \pm 0{,}07$	$2{,}48 \pm 0{,}05$	$2{,}34 \pm 0{,}06$	$1{,}93 \pm 0{,}07$	$1{,}36 \pm 0{,}06$
Octagon [s]	$2{,}40 \pm 1{,}73$	$2{,}41 \pm 0{,}11$	$3{,}05 \pm 0{,}06$	$3{,}34 \pm 0{,}06$	$2{,}96 \pm 0{,}03$	$2{,}43 \pm 0{,}04$	$1{,}69 \pm 0{,}04$
Classroom [s]	$1{,}80 \pm 1{,}12$	$2{,}19 \pm 0{,}11$	$2{,}07 \pm 0{,}05$	$1{,}88 \pm 0{,}03$	$1{,}99 \pm 0{,}02$	$1{,}74 \pm 0{,}02$	$1{,}29 \pm 0{,}01$

En las Figuras 11, 12 y 13 se presentan los esquemas de medición de las salas utilizadas para generar la base de datos.

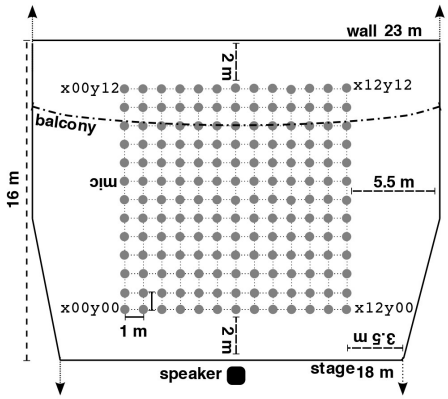

Figura 11. Esquema de medición de respuestas al impulso del recinto Great Hall.

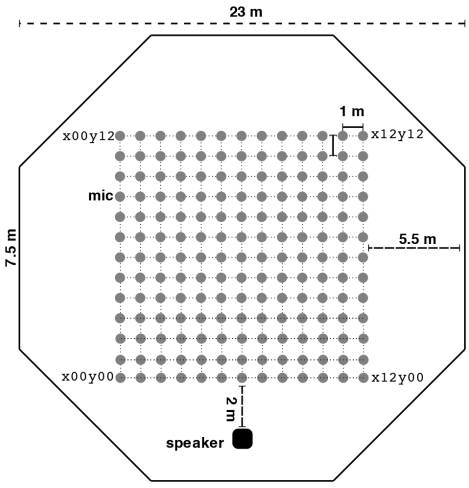

Figura 12. Esquema de medición de respuestas al impulso del recinto Octagon.

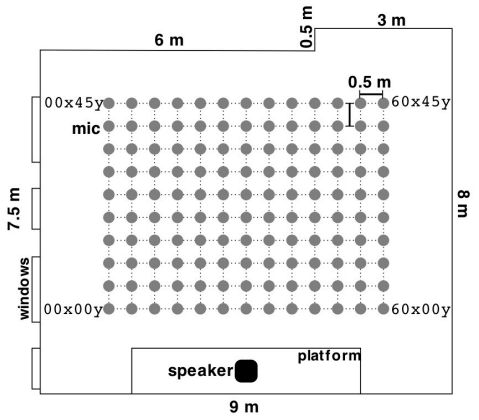

Figura 13. Esquema de medición de respuestas al impulso del recinto Classroom.

4.4.3. Respuestas Al Impulso Aumentadas

Finalmente, se generó el conjunto de respuestas al impulso aumentadas mediante el procesamiento de los impulsos reales de la base de datos previamente mencionada.

Para llevar a cabo esta tarea, se seleccionaron aleatoriamente 15 respuestas al impulso de cada sala, lo que totalizó 45 respuestas para procesar. Estas respuestas se sometieron a las técnicas de aumentación descritas en el capítulo 4.2, con el objetivo de variar los parámetros de tiempo de reverberación (T_{30}) y relación directo-reverberado (DRR).

Específicamente, para el parámetro T_{30}, cada respuesta al impulso se procesó para generar nuevas respuestas con valores de T_{30} que van desde 0.2 s hasta 3 s, en incrementos de 0.1 s. Simultáneamente, se llevó a cabo un proceso adicional de aumentación para el parámetro DRR. Se seleccionaron aleatoriamente 5 de las respuestas al impulso aumentadas en el paso anterior, y se generaron nuevas respuestas con relaciones directo-reverberado que varían desde -6 dB hasta 18 dB, en incrementos de 1 dB.

En consecuencia, para cada respuesta al impulso seleccionada, se realizaron un total de 29 aumentaciones para variar su tiempo de reverberación y 125 aumentaciones para modificar su relación directo-reverberado. Esto da lugar a un procesamiento total de 6930 respuestas al impulso.

Es importante destacar que algunas de las señales a las que se intentó aumentar el tiempo de reverberación no pudieron ser procesadas debido al nivel de ruido intrínseco presente en ciertas bandas de frecuencia. En estos casos, se descartó la respuesta en esa banda específica de la base de datos.

Para llevar a cabo el análisis por banda y aumentar el tiempo de reverberación, se uti-

lizó el banco de filtros de bandas de octava que se muestra en la Figura 14. Se emplearon filtros pasa banda de tipo Butterworth de orden 4 con frecuencias de corte normalizadas para las bandas entre 125 Hz y 4000 Hz, y se utilizó un filtro pasa altos con las mismas características para la banda de 8000 Hz.

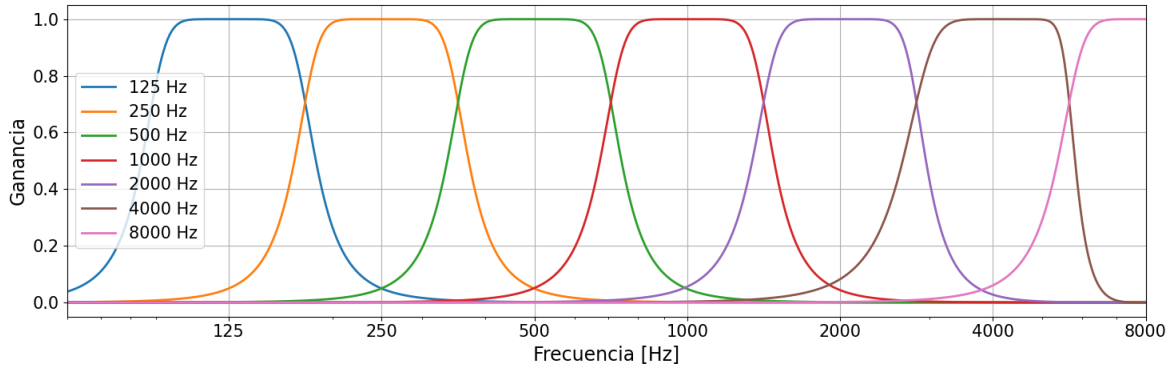

Figura 14. Banco de filtros para el análisis de aumentación de tiempo de reverberación de las respuestas al impulso.

4.5. BASE DE DATOS DE SEÑALES DE VOZ

Como se mencionó previamente, las señales de voz reverberadas se obtienen mediante la convolución de una señal de habla anecoica con respuestas al impulso de salas. Por lo tanto, es necesario contar con una base de datos de señales de voz para llevar a cabo esta investigación.

En este caso, se utilizó la base de datos creada para el desafío *Acoustic Characterisation of Environments Challenge* (ACE Challenge) [15]. Esta base de datos contiene un conjunto de respuestas al impulso grabadas en 5 salas diferentes, así como también grabaciones de voz anecoicas. En cuanto a las señales de habla, se dispone de un total de 50 grabaciones de personas respondiendo preguntas con oraciones extensas, lo que permite generar dinámica en las señales mediante los espacios de silencio entre cada palabra. Para ello, se entrevistó a 5 hombres y 5 mujeres, quienes respondieron a 5 preguntas distintas, dando lugar a los diferentes archivos de audio.

De este conjunto de grabaciones, se descartaron 12 archivos cuya duración era inferior a 5 segundos (el tiempo mínimo requerido para obtener la cantidad de muestras necesarias en las TAE). Luego, se seleccionaron los 38 archivos restantes, dividiéndolos por género y tipo de pregunta. A continuación, se tomaron 30 de estos archivos para conformar la base de datos de entrenamiento, los cuales consistieron en 15 grabaciones de hombres y 15 graba-

ciones de mujeres, respondiendo las mismas preguntas. Los 8 archivos restantes, también divididos equitativamente entre hombres y mujeres, se reservaron para el conjunto de pruebas. Se aseguró que las preguntas fueran diferentes entre el conjunto de entrenamiento y el de pruebas.

Una vez seleccionados los archivos, se recortaron para tener una duración total de 5 segundos y se remuestrearon a una frecuencia de muestreo de 16 kHz.

4.6. CÁLCULO DE LOS DESCRIPTORES DE LAS RESPUESTAS AL IMPULSO

Una vez obtenida la base de datos de respuestas al impulso, el siguiente paso consiste en calcular los descriptores acústicos T_{30}, C_{50}, C_{80} y D_{50} para cada señal. Estos descriptores se calculan siguiendo las directrices establecidas en la norma ISO 3382, como se describió en la sección 3.2.

En todos los casos, las RIRs se filtran utilizando el banco de filtros mostrado en la Figura 14, y luego se calculan los parámetros por bandas de frecuencia.

Dado que algunas bandas pueden presentar niveles de ruido intrínseco elevados en las señales, se optó por estimar el nivel de ruido de fondo utilizando el método de Lundeby al calcular el tiempo de reverberación. Además, se estableció como criterio de aceptación que las bandas tengan un rango dinámico mínimo de 45 dB, desde el punto máximo de la respuesta hasta el nivel estimado del ruido de fondo. En aquellos casos en los que este criterio no se cumple, se descarta esa banda y se continúa con las demás.

Por lo tanto, la cantidad de descriptores disponibles para cada banda de frecuencia se determina exclusivamente por el nivel de ruido de fondo presente en esa banda en particular.

4.7. OBTENCIÓN DE LAS ENVOLVENTES TEMPORALES DE AMPLITUD: TAE

4.7.1. Generación De TAE Sin Ruido

Una vez obtenidas las bases de datos de señales de habla y respuestas al impulso, se procede a convolucionar cada par de datos para obtener las señales reverberadas. A partir de estas señales, se calculan las Envolventes Temporales de Amplitud (TAE).

Para obtener las TAE, se sigue el proceso detallado en la sección 3.5. En primer lugar, se convoluciona la señal de habla con una respuesta al impulso, pero se seleccionan

únicamente los primeros 5 segundos de la señal resultante, en lugar de considerar la convolución completa. Luego, se aplica un filtrado de la señal utilizando el mismo banco de filtros utilizado para la aumentación del tiempo de reverberación, como se muestra en la Figura 14. De esta manera, se obtienen 7 versiones de la señal, una por cada banda de frecuencia correspondiente.

A cada una de estas señales filtradas se le aplica la transformada de Hilbert para obtener la magnitud de su envolvente temporal, y posteriormente se aplica un filtro pasa bajos con una frecuencia de corte de 20 Hz para suavizar aún más la curva. A continuación, se remuestrea cada señal a una frecuencia de muestreo de 40 Hz, con el fin de reducir la tasa de información y facilitar el proceso de entrenamiento de la red neuronal. Finalmente, se normaliza cada señal para que los valores de amplitud se encuentren en el rango de 0 a 1.

Mediante este proceso, se logra una considerable compresión de la información, con una tasa de reducción del 400%. Es importante destacar que se pierde por completo el contenido específico de las palabras, quedando solamente la información relacionada con su dinámica, como se puede observar en la Figura 3.

4.7.2. Generación De TAE Con Ruido Rosa

Para lograr que la red neuronal pueda extraer los parámetros acústicos de las TAE, es importante que estas presenten una amplia dinámica, ya que la estimación se basa en las curvas de decaimiento presentes en el audio. En el caso de las grabaciones anecoicas, esto es sencillo, ya que solo es necesario asegurarse de que el nivel de ruido de la sala de grabación sea bajo. Sin embargo, esto no refleja las condiciones reales a las que se desea aplicar el modelo. En una grabación de audio reverberado en una sala real, es común que exista cierto nivel de ruido propio del entorno, el cual no siempre es posible minimizar y puede ser considerablemente alto, incluso superando los valores propuestos por las normas para la medición de respuestas al impulso.

Los ruidos monótonos y constantes pueden afectar la dinámica de la envolvente temporal de amplitud. Por ejemplo, el ruido producido por las aspas de un ventilador, el murmullo de la gente o el flujo continuo de vehículos. Sin embargo, otros tipos de ruido, como los impulsos, no afectan el análisis de la TAE, ya que durante el proceso de obtención, se reflejan como fluctuaciones en los niveles, lo cual contribuye a la dinámica del audio. Algunos

ejemplos de ruidos impulsivos podrían ser el sonido de una puerta golpeando o la caída de un objeto.

Debido a estas consideraciones, se ha decidido generar una segunda base de datos de envolventes temporales de amplitud agregando ruido constante a las señales. El objetivo es comparar las predicciones de los modelos entrenados utilizando los diferentes conjuntos de datos.

El proceso de obtención de las TAEs con ruido rosa es similar al proceso descrito anteriormente. La diferencia radica en que se agrega ruido controlado a las señales reverberadas para lograr un nivel deseado de relación señal-ruido (SNR) antes de determinar su TAE.

Como se mencionó previamente, los ruidos que presentan mayores desafíos para la tarea de estimación son aquellos monótonos y constantes. Por lo tanto, se decidió agregar ruido rosa a las señales, ya que este tipo de ruido tiende a ser más desafiante para el algoritmo.

El ruido rosa se agrega de manera aleatoria siguiendo una distribución uniforme, con el objetivo de obtener un SNR entre -5 y 20 dB. Para lograr esto, es necesario conocer los valores de las medias cuadráticas (RMS) de la señal reverberada y del ruido. El parámetro SNR se define mediante la ecuación 18.

$$SNR = 10Log_{10}\left(\frac{RMS^2_{señal}}{RMS^2_{ruido}}\right) \tag{18}$$

Dado que conocemos el valor RMS de la señal y del ruido, es posible ajustar la amplitud del ruido multiplicándolo por un factor escalar para obtener el SNR deseado. Este factor se puede obtener utilizando la ecuación 19.

$$a = \frac{\sqrt{\frac{RMS^2_{señal}}{10^{\frac{SNR_{requerido}}{10}}}}}{RMS_{ruido}} \tag{19}$$

Donde a es el factor de compensación de la señal de ruido, $SNR_{requerido}$ es el valor de SNR deseado, RMS_{ruido} es el valor RMS del ruido sin compensar, y $RMS_{señal}$ es el valor RMS de la señal de voz.

Finalmente, la señal reverberada con el nivel de SNR deseado se obtiene mediante la ecuación 20.

$$señal\ con\ ruido = señal + a * ruido \tag{20}$$

4.8. MODELO PROPUESTO

En esta investigación, se propone un sistema para abordar la tarea de estimación de los descriptores acústicos T_{30}, C_{50}, C_{80} y D_{50}. El modelo se basa en el trabajo previo realizado por Duangpummet et al. [11] y se presenta en la Figura 15.

El sistema toma como entrada una señal de habla reverberada con una duración de 5 segundos y una frecuencia de muestreo de 16 kHz. Esta señal se somete a un banco de filtros, como se ilustra en la Figura 14, lo que resulta en la obtención de 7 versiones filtradas de la señal, una para cada banda de frecuencia. En esta etapa, el proceso se paraleliza, ya que cada banda se analiza por subsistemas idénticos.

El primer paso de cada subsistema consiste en calcular las envolventes temporales de amplitud (TAE) utilizando el método descrito en la sección 4.7.1. Estas TAEs se utilizan como parámetros de entrada para una red neuronal convolucional (CNN). La CNN se entrena para estimar los valores de los descriptores acústicos T_{30}, C_{50}, C_{80} y D_{50} a partir de las TAEs. Cabe destacar que se emplean 7 redes neuronales convolucionales, todas con la misma arquitectura, pero entrenadas con valores de TAE filtrados por diferentes bandas de frecuencia.

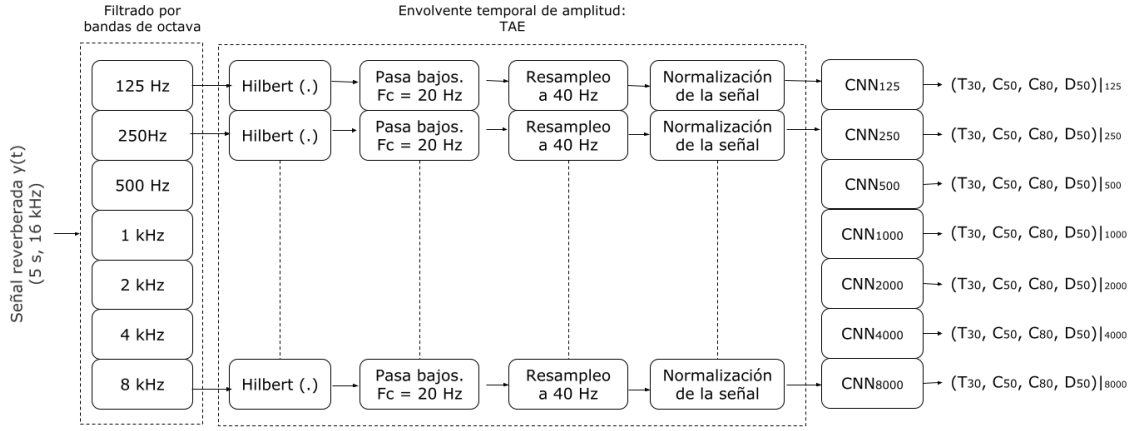

Figura 15. Diagrama en bloques del modelo propuesto.

Cada red neuronal consta de 4 capas convolucionales. La primera capa, llamada capa de entrada, recibe las TAEs como datos de entrada y las convoluciona con filtros. En cada capa convolucional, se aplica una función de activación ReLU para introducir la no linealidad

en la red. Después de cada capa convolucional, se realiza un max pooling para reducir la dimensionalidad, seguido de una normalización por lotes (batch normalization). Se establece una tasa de abandono del 40 % antes de la última capa para evitar el sobreajuste de la red. La capa de salida es una capa completamente conectada que estima los descriptores acústicos utilizando el conocimiento del valor esperado, y se utiliza el error cuadrático medio (MSE) como función de costo para medir la discrepancia entre el valor obtenido y el valor esperado. El modelo se compila utilizando el optimizador Adam [34] con una tasa de aprendizaje de 0.001.

En la tabla 2, se presenta la arquitectura detallada de la red neuronal propuesta para este estudio.

Tabla 2. Arquitectura de red propuesta.

N^o	Tipo de capa	Parámetros
1	Entrada	TAE, tamaño: 200x1
2	$Conv1D^{1era}$	32 filtros, kernel=10, ReLu
3	Pooling	max pooling, pool_size=2,
4	Normalización	batch normalization
5	$Conv1D^{2da}$	18 filtros, kernel=5, ReLu
6	Pooling	max pooling, pool_size=2
7	Normalización	batch normalization
8	Dropout	40 %
9	$Conv1D^{3ra}$	8 filtros, kernel=5, ReLu
10	Pooling	max pooling, pool_size=2
11	Normalización	batch normalization
12	$Conv1D^{4ta}$	4 filtros, kernel=5, ReLu
13	Completamente conectada	Salida 4x1 $(T_{30}, C_{50}, C_{80}, D_{50})$, ReLu
14	Regresión de salida	Error cuadrático medio (MSE)

Se entrenaron dos sistemas utilizando la misma arquitectura de redes neuronales, diferenciándose en los pares de datos de entrada/salida utilizados durante el entrenamiento. En el primer sistema, se emplearon las envolventes temporales de amplitud sin ruido agregado como datos de entrada, mientras que en el segundo sistema se utilizó la base de datos con ruido rosa añadido. En ambos casos, los datos de salida corresponden a los parámetros acústicos que se intentan estimar. La razón detrás de la generación de estas dos versiones se explica detalladamente en la sección 4.7.

En total, se entrenaron 7 redes neuronales para la estimación por bandas de los parámetros acústicos a partir de las señales de habla reverberadas, sin ruido agregado, y otras 7

redes para la estimación utilizando señales con ruido añadido. Todas estas redes poseen la misma arquitectura, variando únicamente los datos de entrada y salida en cada caso.

Durante la etapa de entrenamiento, se realizaron un total de 500 épocas para cada red neuronal, utilizando el error cuadrático medio como función de costo. Esta elección se basó en el trabajo de referencia de Duangpummet et al. [11] y se consideró suficiente para lograr la convergencia del error. Además, se asignó el 70% del conjunto de datos para el entrenamiento, el 10% para la validación y el 20% restante para realizar las estimaciones.

La implementación de las arquitecturas de las redes neuronales se llevó a cabo utilizando la biblioteca Tensorflow [35] en el lenguaje de programación Python.

4.9. EVALUACIÓN DEL MODELO

La evaluación del desempeño de las redes neuronales se lleva a cabo comparando las predicciones de los parámetros acústicos con los valores reales, utilizando tanto una TAE sin ruido (para el primer modelo) como una TAE con ruido rosa agregado (para el segundo modelo).

Este proceso se repite para cada modelo utilizando el conjunto completo de datos de prueba y se calculan los coeficientes de correlación de Spearman [36] para determinar la eficacia de las redes en la estimación de los valores. Estas evaluaciones se realizan para cada una de las 7 bandas de estudio.

Al comparar los valores de correlación obtenidos en cada modelo, se busca determinar si la presencia de ruido en las señales afecta el entrenamiento del sistema y la capacidad para estimar los parámetros acústicos. Esta comparación permitirá evaluar el impacto del ruido en el rendimiento del modelo propuesto.

4.10. ESTUDIO DEL USO DE TÉCNICAS DE AUMENTACIÓN Y SÍNTESIS DE RESPUESTAS AL IMPULSO PARA EL ENTRENAMIENTO DE LA RED

En esta investigación se exploran técnicas de aumentación y síntesis de respuestas al impulso para mejorar el entrenamiento de la red neuronal y su capacidad para estimar los parámetros acústicos. El objetivo es ampliar la base de datos de entrenamiento con una mayor variedad de tiempos de reverberación, lo que se espera que ayude a mejorar la generalización de la estimación en situaciones de medición reales.

Se entrenaron dos modelos de redes neuronales para evaluar si el uso de estas técnicas genera mejoras en el rendimiento del sistema. En el primer modelo, se utilizan solo respuestas al impulso reales, mientras que en el segundo modelo se combina un tercio de respuestas reales, un tercio de respuestas aumentadas y un tercio de respuestas sintéticas. En ambos casos, los archivos de entrenamiento no tienen ruido agregado.

Para determinar qué modelo produce estimaciones más precisas, se calculan los coeficientes de correlación de Spearman entre los valores estimados y los valores reales. Esto se realiza utilizando una base de datos de prueba consistente únicamente en respuestas al impulso reales.

Como prueba adicional, ambos modelos se utilizan para estimar el tiempo de reverberación de una señal de voz grabada en una sala real. Las estimaciones se comparan con el valor real del tiempo de reverberación obtenido mediante el método del barrido frecuencial.

En total, se entrenaron 6 redes neuronales para la estimación de los parámetros acústicos por banda utilizando respuestas al impulso reales, y otras 6 utilizando la combinación de tres tipos de respuestas al impulso. Estas redes abarcan las bandas de frecuencia de 250 Hz a 8000 Hz, excluyendo la banda de 125 Hz debido a la presencia de mucho ruido en los archivos de respuestas al impulso reales.

Cada modelo se entrenó durante 500 épocas utilizando el error cuadrático medio como función de costo.

4.11. COMPARACIÓN DEL MODELO CON MEDICIONES DE CAMPO

Para realizar una comparación adicional, se llevaron a cabo mediciones de campo de respuestas al impulso en tres salas siguiendo las pautas establecidas por la norma ISO 3382 y el método del barrido frecuencial. En cada sala, se seleccionó una posición para la fuente de sonido y dos posiciones para los micrófonos, con el fin de comparar los resultados de los parámetros acústicos obtenidos con diferentes métodos.

Usando el mismo equipo de medición y las mismas configuraciones de micrófonos y fuente de sonido, se grabaron audios de habla de 5 segundos de duración reproducidos a través de un altavoz. Esto permitió capturar la reverberación característica de cada sala en las grabaciones. Se optó por utilizar archivos pregrabados en lugar de realizar la medición en tiempo real con una persona hablando para garantizar que la posición de la fuente de sonido

fuera idéntica en ambos métodos. Además, la reproducción mediante altavoces permitió controlar el nivel de la señal de audio.

Las grabaciones de voz fueron procesadas para obtener sus respectivas envolventes temporales de amplitud (TAE) y luego se pasaron por la red neuronal entrenada con la base de datos sin ruido para estimar los parámetros acústicos de cada sala. Estos resultados se compararon con los obtenidos a través del método normado de las respuestas al impulso.

Para realizar las mediciones, se utilizaron dos micrófonos Behringer ECM 8000 [37] y un altavoz Sony MHC-EC99 [38], que generó un barrido frecuencial de 20 Hz a 13 kHz para obtener las respuestas al impulso de cada sala. Se seleccionó este rango de frecuencias para cubrir las octavas de interés en el estudio, que van desde 125 Hz a 8 kHz. En la Figura 16, se muestra una imagen de una de las salas medidas junto con los instrumentos utilizados.

Figura 16. Imagen de una de las salas medidas.

Por último, en las Figuras 17, 18 y 19 se presentan los esquemas de las tres salas que fueron utilizadas en las mediciones.

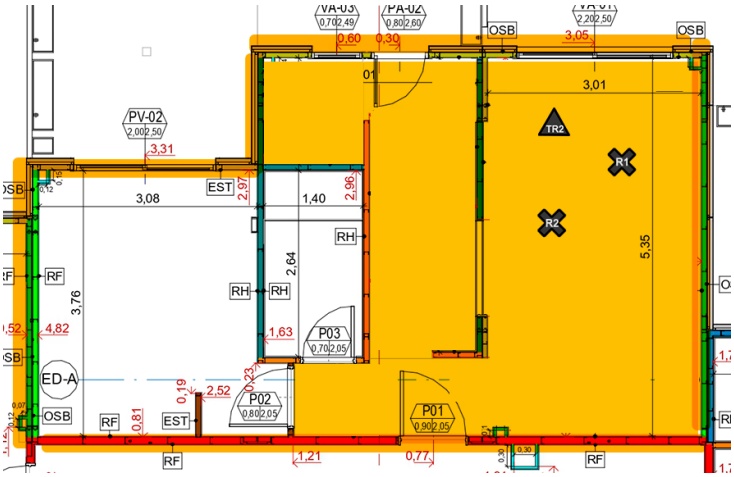

Figura 17. Esquema de medición de la sala 1.

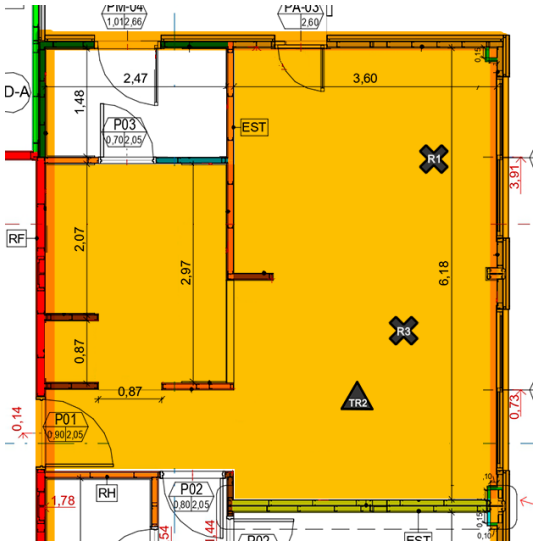

Figura 18. Esquema de medición de la sala 2.

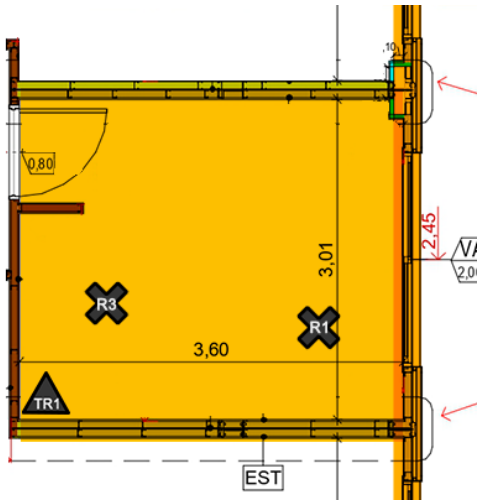

Figura 19. Esquema de medición de la sala 3.

5. RESULTADOS Y ANÁLISIS

En esta sección, se presentarán los resultados obtenidos a partir del modelo propuesto y se realizará un análisis exhaustivo de los mismos. Es importante destacar que el código desarrollado para esta investigación, junto con su correspondiente documentación, se encuentra disponible en un repositorio público en GitHub [39]. Este repositorio incluye las implementaciones de las redes neuronales, los algoritmos utilizados para el cálculo de los descriptores acústicos, así como el análisis de los datos obtenidos.

Inicialmente, se tenía la intención de realizar una comparación con los resultados obtenidos en el trabajo de Duangpummet et al. [11]. Sin embargo, esta comparación no fue posible debido a que la base de datos utilizada en dicho trabajo no se encuentra públicamente disponible.

5.1. ANÁLISIS DE LA BASE DE DATOS DE RESPUESTAS AL IMPULSO

Como se mencionó anteriormente, para que la red neuronal pueda generalizar el aprendizaje y estimar con precisión los parámetros acústicos de salas desconocidas, es fundamental contar con una base de datos que represente una amplia variedad de casos encontrados en mediciones reales.

En el contexto de esta investigación, se buscó tener una representación diversa de los recintos. Para lograr esto, se analizaron los tiempos de reverberación y las relaciones directo-reverberado en las respuestas al impulso (RIRs) de la base de datos.

Se eligieron dos descriptores acústicos: el tiempo de reverberación medio ($T30_{mid}$) y la relación directo-reverberado media (DRR_{mid}) en las bandas de 500 y 1000 Hz.

En la Figura 20, se muestran los valores de estos descriptores para las RIRs generadas de forma sintética. Este análisis permite examinar la distribución y la variabilidad de los tiempos de reverberación y las relaciones directo-reverberado en la base de datos.

En este conjunto de datos, se observa una disminución en el valor del DRR a medida que aumenta el tiempo de reverberación en las señales. Este resultado concuerda con las expectativas teóricas, ya que indica que es difícil obtener valores altos de DRR junto con tiempos de reverberación largos. Esto se debe a que un mayor tiempo de reverberación implica una mayor cantidad de energía en la parte correspondiente al campo reverberado

dentro de la respuesta al impulso (RIR). Por lo tanto, utilizando únicamente esta técnica de síntesis, no es posible simular la totalidad de recintos que se encuentran en mediciones reales.

Para abordar esta limitación, sería necesario utilizar alguna otra técnica que permita manipular el valor del DRR sin afectar el tiempo de reverberación durante la síntesis de las RIR.

El segundo grupo a analizar corresponde a las respuestas reales de la base de datos, las cuales se presentan en la Figura 21.

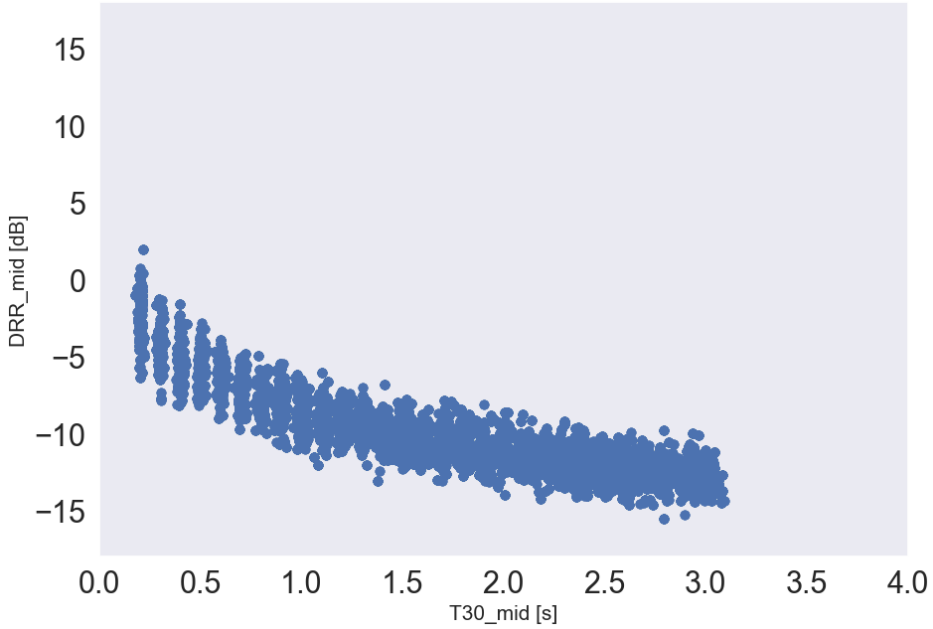

Figura 20. Gráfico de dispersión del DRR_{mid} contra el $T30_{mid}$ para las RIR sintéticas.

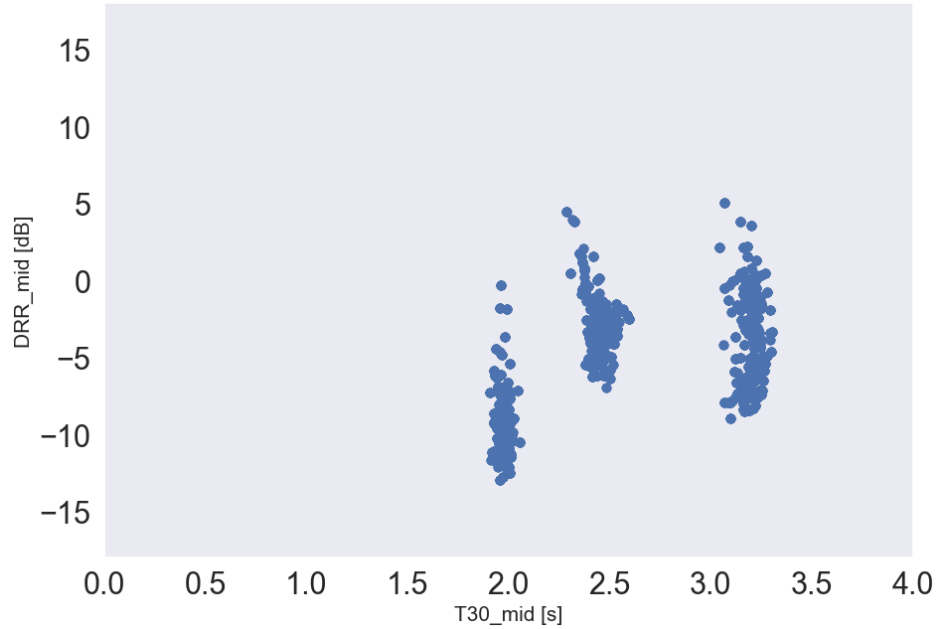

Figura 21. Gráfico de dispersión del DRR_{mid} contra el $T30_{mid}$ para las RIR de salas reales.

El gráfico muestra claramente los tres recintos medidos, con el primer grupo de puntos correspondiente al *classroom*, seguido por el *great hall* y, finalmente, el *octagon* con el tiempo de reverberación más largo.

Se puede observar que la base de datos de salas reales presenta una considerable variación en el valor del DRR, debido a la cantidad de puntos de medición tomados en cada recinto. Sin embargo, esta variación no se refleja en el tiempo de reverberación. Dado que estas variaciones en el DRR están asociadas a un número limitado de valores de T_{30}, se resalta la necesidad de utilizar técnicas de aumentación para representar una mayor diversidad de recintos.

La Figura 22 muestra los valores de $T30_{mid}$ y DRR_{mid} obtenidos después de procesar las respuestas reales mediante técnicas de aumentación.

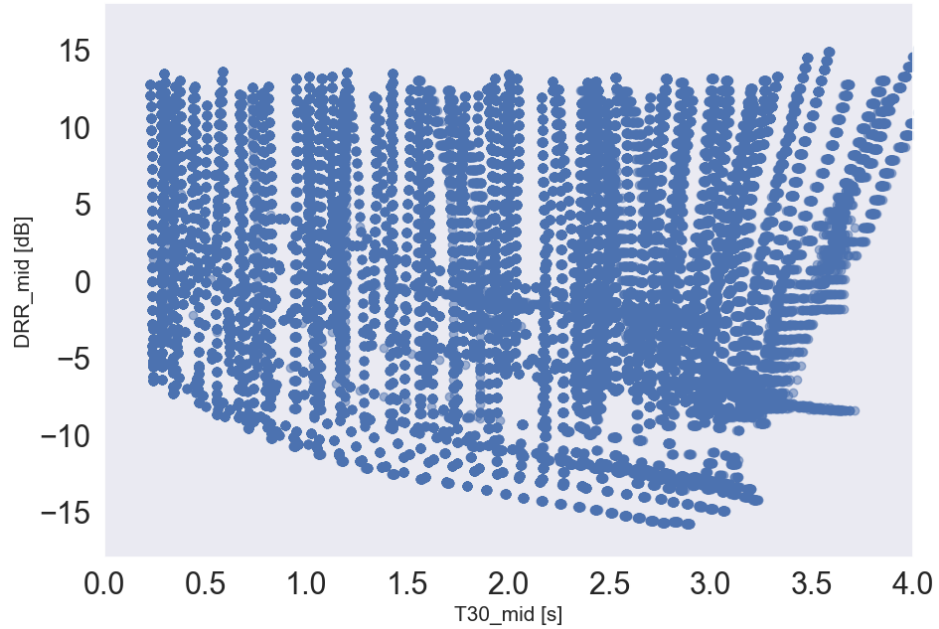

Figura 22. Gráfico de dispersión del DRR_{mid} contra el $T30_{mid}$ para las RIR aumentadas.

En este último conjunto de respuestas se observa una mayor diversidad de recintos simulados en comparación con los dos conjuntos anteriores. Además, este análisis valida el funcionamiento de los algoritmos de aumentación tanto para el T_{30} como para el DRR.

Al combinar los grupos de respuestas reales, sintéticas y aumentadas, se logra formar una base de datos más completa que representa una amplia variedad de recintos. En esta base de datos, los valores de $T30_{mid}$ varían desde 0.18 s hasta 4.09 s, mientras que los valores de DRR_{mid} oscilan entre -15.78 dB y 16.77 dB.

Para tener una mejor comprensión global de la base de datos, se presentan diagramas de cajas en las Figuras 23 y 24, que muestran los valores de los descriptores T_{30} y DRR para todas las bandas de frecuencia analizadas.

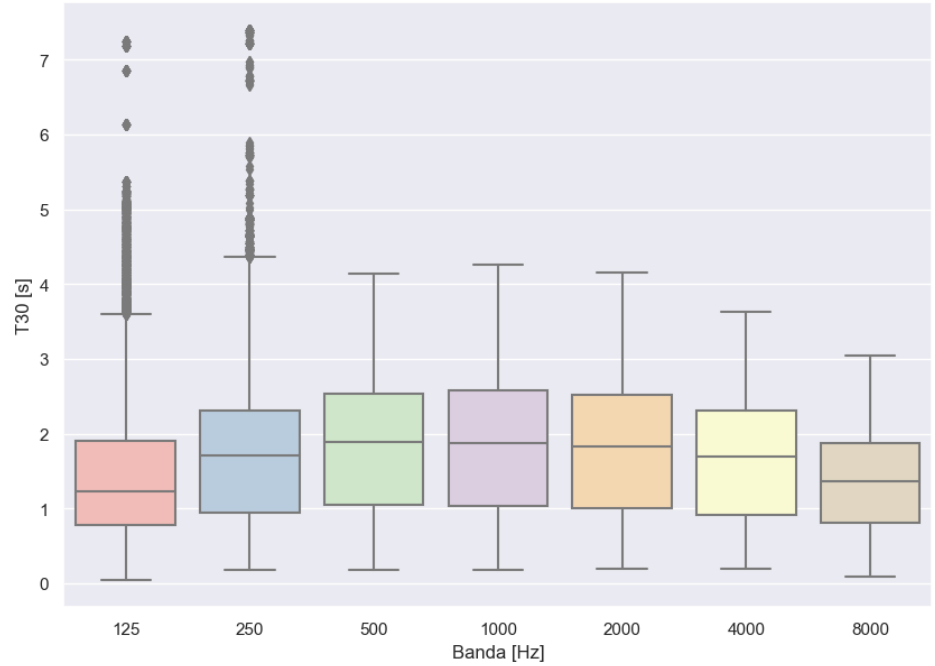

Figura 23. Boxplot de los valores de T_{30} por bandas de octava de toda la base de datos.

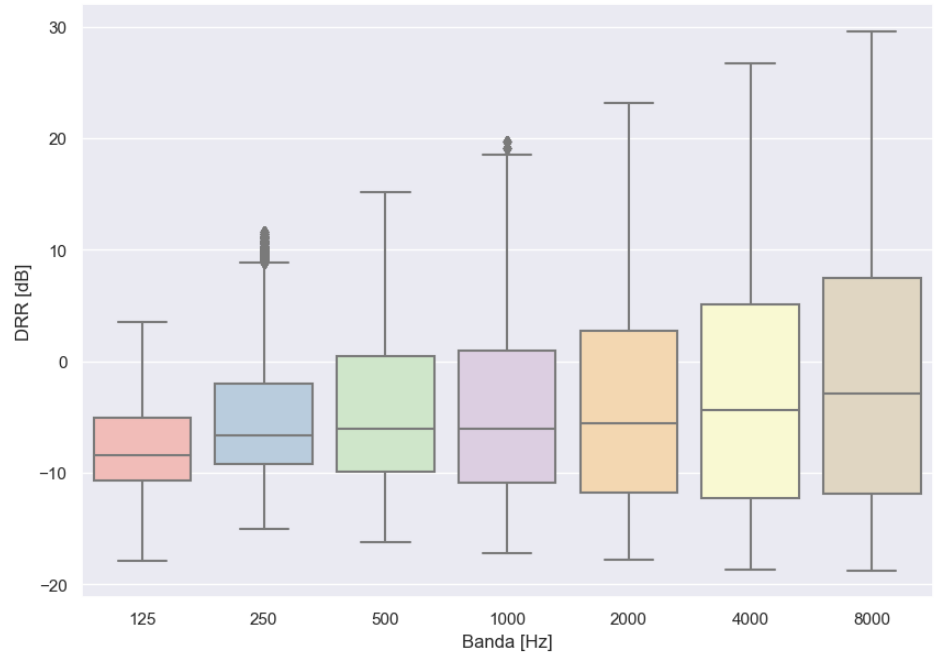

Figura 24. Boxplot de los valores de DRR por bandas de octava de toda la base de datos.

Durante el cálculo del descriptor T_{30} para las bandas de 125 Hz y 250 Hz, se observaron numerosos valores atípicos debido al ruido inherente en esas frecuencias. Este ruido provoca un aumento en la pendiente de los cuadrados mínimos y, por lo tanto, en el valor final del parámetro. A pesar de esto, se decidió mantener estos valores en la base de datos para

evaluar el rendimiento del algoritmo en presencia de estos casos atípicos. Se espera que el algoritmo funcione mejor en frecuencias más altas, según el análisis preliminar de estos datos.

Por otro lado, en la tabla 3 se muestra el número total de respuestas al impulso por banda que se conservaron en la base de datos después del análisis y cálculo de los descriptores. También se indica el total de horas de entrenamiento disponibles para cada banda.

Tabla 3. Cantidad de RIRs en la base de datos y total de duración en horas de la misma por bandas de frecuencia.

	Banda [Hz]						
	125	250	500	1000	2000	4000	8000
Total de RIRs [n]	3768	3929	4135	4141	4141	4145	4145
Duración de la BD [h]	159.1	165.9	174.6	174.8	174.9	175.0	175.0

Se observó que el número de respuestas al impulso aumenta a medida que se incrementa la banda de frecuencia analizada. Este aumento se debe a que en las bandas de frecuencias más bajas, las respuestas al impulso suelen contener ruido que dificulta el cálculo preciso de los descriptores. Por lo tanto, se descartan las respuestas al impulso en esas bandas, lo que resulta en una menor cantidad de archivos conservados en las frecuencias más bajas. Este hecho sugiere que el algoritmo puede tener un mejor rendimiento en frecuencias altas, ya que cuenta con más datos para su entrenamiento en esas bandas.

5.2. ENTRENAMIENTO DE LOS MODELOS

Una vez generadas y analizadas las bases de datos, se procedió al entrenamiento de los modelos de acuerdo a lo establecido en la sección 4.8.

En total, se entrenaron 7 redes neuronales para la estimación por bandas de los parámetros acústicos utilizando señales de habla reverberadas sin ruido, y otras 7 redes utilizando señales con ruido agregado. Todas las redes neuronales tenían la misma arquitectura, diferenciándose únicamente en los datos de entrada y salida. Siguiendo el trabajo de referencia de Duangpummet et al. [11], se utilizó un total de 500 épocas durante la etapa de entrenamiento, utilizando el error cuadrático medio como función de costo.

La arquitectura descrita en la tabla 2 y la elección de 500 épocas permitieron al algoritmo reducir la función de costo durante el entrenamiento sin llegar a memorizar los datos. Esto se puede observar en la Figura 25, donde tanto la función de costo como la de validación

disminuyen sin que la de validación comience a aumentar, lo cual indicaría un sobreajuste o overfitting del modelo.

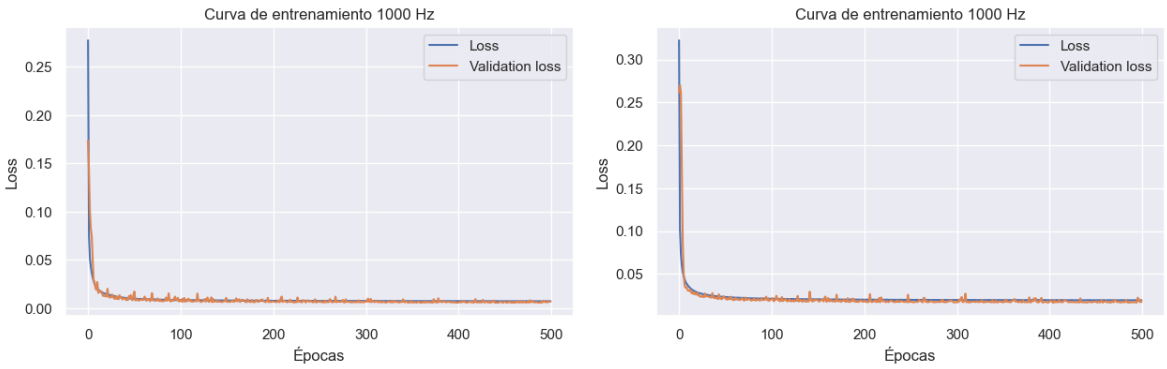

Figura 25. Curvas de entrenamiento para la banda de 1000 Hz de los modelos sin ruido (izquierda) y con ruido (derecha) respectivamente.

Basado en los resultados del entrenamiento, se pudo observar que el modelo propuesto es capaz de estimar los descriptores acústicos a partir de la TAE de un audio de voz reverberado, tanto en presencia como en ausencia de ruido. Este comportamiento se evidenció en todas las bandas de frecuencia estudiadas.

Aunque en ambos casos los valores de la función de costo se minimizan, para detectar posibles diferencias significativas en la convergencia de los errores cuadráticos medios, es necesario examinar cada valor individualmente. En la tabla 4 se presentan los valores de la función de costo obtenidos en la última época de entrenamiento para cada banda de frecuencia, tanto en los modelos entrenados con TAEs sin ruido como en los entrenados con ruido agregado.

Tabla 4. Valores de error cuadrático medio (ECM) en el set de entrenamiento y de validación obtenidos durante el entrenamiento de las redes neuronales usando la base de datos con ruido y sin ruido

Banda [Hz]	Base de datos sin ruido		Base de datos con ruido	
	ECM entrenamiento	ECM validación	ECM entrenamiento	ECM validación
125	0.068	0.053	0.082	0.070
250	0.029	0.035	0.044	0.039
500	0.012	0.012	0.030	0.027
1000	0.007	0.007	0.019	0.017
2000	0.005	0.005	0.016	0.015
4000	0.003	0.002	0.014	0.013
8000	0.003	0.002	0.012	0.011

Como se pudo intuir al observar el diagrama de cajas en la Figura 23, la banda de

frecuencia con el mayor error cuadrático medio es la de 125 Hz. Esto se debe principalmente a la presencia de valores atípicos en la base de datos en esta frecuencia, así como a la menor cantidad de horas de entrenamiento disponibles para esta banda. Además, se obtuvieron resultados significativamente mejores para la base de datos sin ruido en comparación con la base de datos con ruido rosa agregado. Esto confirma la hipótesis de que la adición de ruido introduce una mayor complejidad en la tarea de estimación para la red neuronal. Este comportamiento era esperado, ya que la adición de ruido disminuye el rango dinámico del audio y, por lo tanto, también afecta la dinámica de las pendientes de decaimiento en la TAE. Este fenómeno se asemeja a lo que ocurre cuando se calcula la RIR de una sala y la presencia de ruido de fondo durante la medición afecta la precisión en la estimación del tiempo de reverberación.

5.3. ANÁLISIS DEL MODELO ENTRENADO SOLO CON RESPUESTAS REALES CONTRA RESPUESTAS COMBINADAS

En esta sección, analizamos la hipótesis de que el uso de técnicas de aumentación de respuestas al impulso mejora el entrenamiento y la estimación de parámetros acústicos en comparación con mediciones realizadas en salas reales.

Para esto, se entrenaron dos sistemas, cada uno compuesto por 6 redes neuronales. La única diferencia entre ellos radica en que uno utilizó únicamente respuestas al impulso reales para generar las TAE de entrenamiento, mientras que el otro utilizó una combinación equitativa de respuestas reales, sintéticas y aumentadas, de modo que ambas bases de datos generadas tuvieran la misma cantidad de archivos.

Una vez entrenados los sistemas, se realizaron estimaciones utilizando una base de TAE generadas solo con respuestas al impulso reales y audios de voz. A partir de estos resultados, se calcularon los coeficientes de correlación de Spearman entre la diferencia entre el valor estimado y el valor esperado. Esto se realizó para cada banda de estudio y para los descriptores T_{30}, C_{50}, C_{80} y D_{50}.

Los resultados obtenidos para cada descriptor se presentan en las tablas 5, 6, 7 y 8.

Tabla 5. Coeficiente de correlación de Spearman entre las predicciones y el valor real de T_{30} por cada banda y por base de datos utilizada.

Banda [Hz]	Base de datos con RIRs reales		Base de datos combinada	
	r	Valor p	r	Valor p
250	0.243	<0.01	0.887	<0.01
500	0.304	<0.01	0.945	<0.01
1000	0.311	<0.01	0.945	<0.01
2000	0.321	<0.01	0.955	<0.01
4000	0.761	<0.01	0.960	<0.01
8000	0.780	<0.01	0.920	<0.01

Tabla 6. Coeficiente de correlación de Spearman entre las predicciones y el valor real de C_{50} por cada banda y por base de datos utilizada.

Banda [Hz]	Base de datos con RIRs reales		Base de datos combinada	
	r	Valor p	r	Valor p
250	0.786	<0.01	0.602	<0.01
500	0.813	<0.01	0.849	<0.01
1000	0.762	<0.01	0.894	<0.01
2000	0.850	<0.01	0.900	<0.01
4000	0.913	<0.01	0.962	<0.01
8000	0.955	<0.01	0.974	<0.01

Tabla 7. Coeficiente de correlación de Spearman entre las predicciones y el valor real de C_{80} por cada banda y por base de datos utilizada.

Banda [Hz]	Base de datos con RIRs reales		Base de datos combinada	
	r	Valor p	r	Valor p
250	0.770	<0.01	0.841	<0.01
500	0.829	<0.01	0.919	<0.01
1000	0.700	<0.01	0.930	<0.01
2000	0.814	<0.01	0.947	<0.01
4000	0.899	<0.01	0.974	<0.01
8000	0.942	<0.01	0.981	<0.01

Tabla 8. Coeficiente de correlación de Spearman entre las predicciones y el valor real de D_{50} por cada banda y por base de datos utilizada.

Banda [Hz]	Base de datos con RIRs reales		Base de datos combinada	
	r	Valor p	r	Valor p
250	0.779	<0.01	0.621	<0.01
500	0.829	<0.01	0.844	<0.01
1000	0.742	<0.01	0.892	<0.01
2000	0.840	<0.01	0.899	<0.01
4000	0.922	<0.01	0.961	<0.01
8000	0.954	<0.01	0.964	<0.01

Los resultados obtenidos en las tablas anteriores demuestran que el agregado de respuestas al impulso aumentadas y sintéticas ayudó a mejorar el entrenamiento de la red para estimar los parámetros acústicos, especialmente en el caso del tiempo de reverberación (T_{30}). Esto se debe a que el sistema aprendió a calcular los descriptores en un rango más amplio de valores, lo que resulta en una menor desviación al presentarle un valor desconocido durante la fase de prueba.

La mejora en la estimación es más notable en el caso del tiempo de reverberación debido a la limitada variabilidad de la base de datos de respuestas reales, como se observó en la Figura 21. La red tenderá a estimar los valores que conoce, lo que se refleja en una menor capacidad para generalizar a valores fuera de ese rango. Sin embargo, al aumentar la diversidad de los datos de entrenamiento, se logra mejorar significativamente este aspecto.

Como segunda prueba, se compararon ambos sistemas utilizando una grabación de voz in situ realizada en una sala para esta investigación. Los resultados de esta estimación se compararon con el tiempo de reverberación obtenido a través de la técnica del LSS, asegurando que tanto la fuente como los micrófonos estuvieran en la misma posición durante la grabación del audio de voz y el barrido de frecuencia. Los resultados de esta comparación se presentan en la tabla 9.

De acuerdo con los resultados de la tabla 9, se puede observar que el sistema entrenado con la base de datos combinada de respuestas al impulso generó mejores estimaciones en comparación con el sistema entrenado solo con respuestas reales. En este caso, la sala medida tiene un tiempo de reverberación cercano a $1\,s$, el cual fue estimado con cierta precisión por el sistema entrenado con la base de datos combinada. Por otro lado, el sistema entrenado solo con respuestas reales tendió a generar estimaciones por encima de los $2\,s$, que son los valores que predominan en las respuestas al impulso utilizadas para su entrenamiento, como se observa en la tabla 9.

Por lo tanto, se puede concluir que el agregado de respuestas sintéticas y aumentadas genera una mejora en la estimación del sistema frente a respuestas reales desconocidas, permitiendo una mayor generalización y adaptabilidad a diferentes condiciones acústicas.

Tabla 9. Comparación del T_{30} obtenido de una medición in situ por la red entrenada con RIRs reales contra la entrenada con todos los tipos de RIRs combinados para una medición

Obtenido de	Banda [Hz]	T_{30} [s]
BD reales		2.261
BD combinadas	250	0.896
RIR		1.269
BD reales		2.526
BD combinadas	500	1.634
RIR		1.645
BD reales		2.249
BD combinadas	1000	1.567
RIR		1.672
BD reales		2.326
BD combinadas	2000	1.463
RIR		1.340
BD reales		1.761
BD combinadas	4000	1.028
RIR		0.954
BD reales		1.206
BD combinadas	8000	1.002
RIR		0.564

5.4. EVALUACIÓN DE LOS MODELOS CON AUDIOS DESCONOCIDOS

Se procedió a evaluar la eficacia de los modelos en la predicción de los descriptores de la sala utilizando audios desconocidos, después de haber demostrado previamente que el uso de respuestas sintéticas y aumentadas mejora la estimación del algoritmo, siguiendo los lineamientos presentados en la sección 4.8.

En este caso, se calculó el coeficiente de correlación de Spearman r (junto con su valor p asociado) comparando los valores predichos por la red neuronal con los valores calculados a partir de las RIRs. Se optó por utilizar este coeficiente de correlación en lugar del r^2, ya que es el que suele emplearse en trabajos de estimación ciega de parámetros acústicos [11][19].

Dado que durante el entrenamiento y la posterior estimación se utilizaron TAEs generadas a partir de RIRs reales, sintéticas y aumentadas, se decidió calcular los coeficientes de correlación por separado para cada uno de estos grupos y luego compararlos tomando el promedio. El objetivo de este análisis es determinar si la estimación de los descriptores mejora en función del tipo de RIR utilizado para generar las TAEs.

Las estimaciones de los descriptores se agrupan según el tipo de descriptor, la banda

de frecuencia y si se utilizó la red entrenada con una base de datos con ruido o sin ruido. En cada caso, se calcula el coeficiente de correlación promedio ($r_{promedio}$) a partir de las predicciones realizadas sobre TAEs reales (r_{reales}), TAEs sintéticas ($r_{sinteticas}$) y TAEs aumentadas ($r_{aumentadas}$). Sin embargo, en la banda de 125 Hz no se disponen de archivos de TAEs obtenidas a partir de RIRs reales debido al ruido presente en esa banda. En consecuencia, en esta banda se calcula el promedio utilizando los coeficientes de correlación de las TAEs sintéticas y aumentadas ($r_{sinteticas}$ y $r_{aumentadas}$).

En las tablas 10, 11, 12 y 13 se presentan los valores de los coeficientes de correlación obtenidos para los descriptores T_{30}, C_{50}, C_{80} y D_{50}, respectivamente. Estos valores se muestran para cada banda de frecuencia y base de datos utilizada durante el entrenamiento, lo que permite evaluar el rendimiento de los modelos y determinar posibles diferencias significativas en la capacidad de predicción según la base de datos empleada.

Tabla 10. Coeficiente de correlación de Spearman entre las predicciones y el valor real de T_{30} por cada banda y por base de datos utilizada.

Obtenido de	Banda [Hz]	r_{reales}	$r_{sinteticas}$	$r_{aumentadas}$	$r_{promedio}$
BD sin ruido	125	-	0.714	0.718	0.716
BD con ruido		-	0.633	0.645	0.639
BD sin ruido	250	0.934	0.935	0.938	0.936
BD con ruido		0.759	0.761	0.770	0.763
BD sin ruido	500	0.952	0.954	0.953	0.953
BD con ruido		0.725	0.730	0.714	0.723
BD sin ruido	1000	0.959	0.961	0.961	0.960
BD con ruido		0.807	0.821	0.805	0.811
BD sin ruido	2000	0.959	0.956	0.957	0.957
BD con ruido		0.855	0.858	0.849	0.854
BD sin ruido	4000	0.939	0.953	0.943	0.945
BD con ruido		0.885	0.886	0.887	0.886
BD sin ruido	8000	0.953	0.950	0.949	0.951
BD con ruido		0.923	0.908	0.912	0.914

Tabla 11. Coeficiente de correlación de Spearman entre las predicciones y el valor real de C_{50} por cada banda y por base de datos utilizada.

Obtenido de	Banda [Hz]	r_{reales}	$r_{sinteticas}$	$r_{aumentadas}$	$r_{promedio}$
BD sin ruido	125	-	0.649	0.638	0.644
BD con ruido		-	0.669	0.669	0.669
BD sin ruido	250	0.668	0.681	0.666	0.672
BD con ruido		0.661	0.662	0.669	0.664
BD sin ruido	500	0.819	0.837	0.837	0.831
BD con ruido		0.715	0.768	0.755	0.746
BD sin ruido	1000	0.916	0.916	0.907	0.913
BD con ruido		0.814	0.842	0.835	0.830
BD sin ruido	2000	0.927	0.938	0.929	0.931
BD con ruido		0.828	0.837	0.840	0.835
BD sin ruido	4000	0.967	0.970	0.970	0.969
BD con ruido		0.836	0.844	0.845	0.842
BD sin ruido	8000	0.966	0.972	0.969	0.969
BD con ruido		0.847	0.843	0.845	0.845

Tabla 12. Coeficiente de correlación de Spearman entre las predicciones y el valor real de C_{80} por cada banda y por base de datos utilizada.

Obtenido de	Banda [Hz]	r_{reales}	$r_{sinteticas}$	$r_{aumentadas}$	$r_{promedio}$
BD sin ruido	125	-	0.820	0.813	0.817
BD con ruido		-	0.764	0.774	0.769
BD sin ruido	250	0.876	0.872	0.876	0.875
BD con ruido		0.809	0.785	0.795	0.796
BD sin ruido	500	0.903	0.917	0.917	0.912
BD con ruido		0.786	0.826	0.806	0.806
BD sin ruido	1000	0.951	0.954	0.950	0.952
BD con ruido		0.859	0.880	0.872	0.870
BD sin ruido	2000	0.959	0.966	0.961	0.962
BD con ruido		0.878	0.879	0.883	0.880
BD sin ruido	4000	0.981	0.984	0.984	0.983
BD con ruido		0.880	0.885	0.885	0.883
BD sin ruido	8000	0.981	0.982	0.982	0.982
BD con ruido		0.883	0.876	0.878	0.879

Tabla 13. Coeficiente de correlación de Spearman entre las predicciones y el valor real de D_{50} por cada banda y por base de datos utilizada.

Obtenido de	Banda [Hz]	r_{reales}	$r_{sinteticas}$	$r_{aumentadas}$	$r_{promedio}$
BD sin ruido	125	-	0.642	0.637	0.640
BD con ruido		-	0.667	0.664	0.666
BD sin ruido	250	0.675	0.682	0.669	0.675
BD con ruido		0.671	0.667	0.676	0.671
BD sin ruido	500	0.816	0.834	0.836	0.829
BD con ruido		0.717	0.770	0.757	0.748
BD sin ruido	1000	0.909	0.910	0.900	0.906
BD con ruido		0.813	0.843	0.835	0.830
BD sin ruido	2000	0.930	0.938	0.931	0.933
BD con ruido		0.828	0.837	0.841	0.835
BD sin ruido	4000	0.963	0.968	0.967	0.966
BD con ruido		0.844	0.846	0.850	0.847
BD sin ruido	8000	0.965	0.970	0.966	0.967
BD con ruido		0.846	0.839	0.839	0.841

Se observaron diferencias mínimas entre los coeficientes de correlación calculados con las TAEs generadas a partir de RIRs reales, sintéticas y aumentadas. Esto sugiere que los tres tipos de respuestas al impulso son igualmente válidos para generar los conjuntos de datos utilizados en el entrenamiento de la red neuronal, ya que la red no muestra preferencia por alguno de ellos. Por lo tanto, las estimaciones de los descriptores son similares en los tres casos.

Como se sospechaba al analizar los diagramas de caja de las bases de datos, se observó que la banda de 125 Hz presenta la correlación más baja en los cuatro descriptores analizados. Aunque solo se realizó un análisis de valores atípicos en esta banda para el tiempo de reverberación, estos resultados indican que existe cierta correlación entre los elementos que afectan el cálculo de los descriptores. Es decir, la falta de rango dinámico en los audios también afecta la estimación de los demás descriptores, lo cual no es evidente al analizar las fórmulas de cálculo de los mismos.

Por otro lado, como era de esperar, los coeficientes de correlación utilizando la base de datos con ruido rosa agregado fueron inferiores en comparación con el otro conjunto de datos en todos los casos analizados. Sin embargo, todos los valores calculados están por encima del 63%, lo que indica que la red es capaz de estimar los descriptores incluso en presencia de ruido de fondo, aunque con menor precisión.

A excepción de la banda de 125 Hz, se obtuvieron coeficientes de correlación promedio superiores al 93 % para el descriptor T_{30} en la red entrenada sin ruido agregado. Esto demuestra la capacidad de la red para estimar este descriptor con gran precisión en general, a lo largo de las distintas bandas.

En cuanto a los demás descriptores, el segundo en presentar mejores estimaciones fue C_{80}, con valores por encima del 87 %. Le siguió D_{50}, con valores superiores al 67 %. Por último, el descriptor C_{50} que también mostró valores superiores al 67 %. Aunque los coeficientes de correlación obtenidos para los cuatro descriptores no difieren significativamente, todos ellos presentan niveles muy altos. En consecuencia, se puede concluir que la red es capaz de estimar de manera efectiva los cuatro parámetros acústicos en todas las bandas.

Exceptuando la banda de 125 Hz, en el caso particular de los coeficientes de correlación promedio para el descriptor T_{30} en la red entrenada sin ruido agregado, se obtuvieron valores superiores al 93 %, lo que demuestra que la red puede estimar este descriptor con gran precisión.

El segundo descriptor con mejores estimaciones fue C_{80} con valores por encima del 87 %, seguido por D_{50} con valores por encima del 67 % y, por último, C_{50} también con valores por encima del 67 %. Sin embargo, los valores de correlación obtenidos en los cuatro descriptores no difieren significativamente y se consideran muy altos. Por lo tanto, se puede concluir que la red logra estimar los cuatro parámetros acústicos de manera efectiva en todas las bandas.

Aunque se esperaba que las mayores correlaciones se encontraran en las bandas de baja frecuencia (entre 100 y 300 Hz aproximadamente), dado que los descriptores se calcularon a partir de audios de voz, esto no se cumplió para ningún parámetro acústico. No se puede determinar con certeza la razón de este comportamiento, pero se puede concluir que se debe a que el ruido intrínseco del audio es menor en altas frecuencias, lo que resulta en un mayor rango dinámico y un mejor perfil de decaimiento de la curva utilizado en la estimación.

5.5. EVALUACIÓN DEL MODELO CON MEDICIONES REALES

Como último paso, se llevó a cabo la evaluación de los modelos utilizando audios grabados en varias salas reales. En concreto, se realizaron barridos frecuenciales y se grabaron

audios de voz en tres salas diferentes, con dos puntos de medición en cada una.

A partir de los barridos frecuenciales, se calculó la respuesta al impulso de cada sala (RIR), y se obtuvieron los valores de los descriptores utilizando el software REW [40]. La elección de este programa se basó en su reputación en la comunidad científica, su amplia gama de características y su accesibilidad como software gratuito. Paralelamente, los audios de voz se procesaron para obtener las TAEs correspondientes, las cuales se utilizaron para estimar los valores de los descriptores utilizando los modelos entrenados.

Con el fin de asegurar que tanto los audios de voz como los barridos frecuenciales se tomaran desde la misma posición, ambos fueron emitidos utilizando el mismo parlante dentro de cada sala.

Por último, se utilizaron los valores de la diferencia apenas perceptible (JND, por sus siglas en inglés, *just noticeable difference*) de los descriptores como parámetro de referencia para evaluar qué tan cercanas estaban las estimaciones a una variación apenas perceptible. Sin embargo, dado que este parámetro es muy restrictivo, no se consideró como un valor determinante para evaluar la calidad de las estimaciones. Los valores de JND se encuentran en la tabla 14.

Tabla 14. JND de los descriptores acústicos.

Parámetro	JND
T_{30} [s]	5 %
C_{50} [dB]	1 dB
C_{80} [dB]	1 dB
D_{50} [%]	5 %

5.5.1. Sala 1

Para la primera sala seleccionada, se realizaron grabaciones de barridos frecuenciales y audios de voz en dos puntos distintos, y luego se calcularon los resultados promediados.

La tabla 15 muestra los valores estimados por la red neuronal en comparación con los valores obtenidos mediante métodos convencionales a partir de la respuesta al impulso para la posición 1.

Tabla 15. Valores de los descriptores calculados a partir de una respuesta al impulso contra los estimados a partir de la grabación de voz con la red neuronal para la posición 1.

Obtenido de	Banda [Hz]	T_{30} [s]	C_{50} [dB]	C_{80} [dB]	D_{50} [%]
RIR		0.64	-0.33	3.56	49.44
Audio de voz	125	0.71	-0.26	3.87	48.50
JND		0.64 ± 0.03	-0.33 ± 1.00	3.56 ± 1.00	71.21 ± 5.00
RIR		1.13	-2.19	2.63	38.74
Audio de voz	250	1.27	2.22	4.59	62.50
JND		1.13 ± 0.06	-2.19 ± 1.00	2.63 ± 1.00	38.74 ± 5.00
RIR		1.87	-2.89	-0.67	37.81
Audio de voz	500	1.65	-2.68	2.64	35.10
JND		1.87 ± 0.09	-2.89 ± 1.00	-0.67 ± 1.00	37.81 ± 5.00
RIR		1.63	-2.88	0.05	35.32
Audio de voz	1000	1.67	-1.38	0.33	42.10
JND		1.63 ± 0.08	-2.88 ± 1.00	0.05 ± 1.00	35.32 ± 5.00
RIR		1.54	-3.19	-0.31	34.06
Audio de voz	2000	1.34	0.72	4.02	54.10
JND		1.54 ± 0.08	-3.19 ± 1.00	-0.31 ± 1.00	34.06 ± 5.00
RIR		1.08	-1.18	2.11	44.02
Audio de voz	4000	0.95	1.39	4.09	58.00
JND		1.08 ± 0.05	-1.18 ± 1.00	2.11 ± 1.00	44.02 ± 5.00
RIR		1.38	0.45	2.70	51.93
Audio de voz	8000	0.56	4.82	8.50	75.20
JND		1.38 ± 0.07	0.45 ± 1.00	2.70 ± 1.00	51.93 ± 5.00

De manera similar, la tabla 16 presenta la misma información pero para la posición 2.

Tabla 16. Valores de los descriptores calculados a partir de una respuesta al impulso contra los estimados a partir de la grabación de voz con la red neuronal para la posición 2.

Obtenido de	Banda [Hz]	T_{30} [s]	C_{50} [dB]	C_{80} [dB]	D_{50} [%]
RIR		1.09	4.86	7.67	71.21
Audio de voz	125	0.65	6.80	13.41	82.70
JND		1.09 ± 0.05	4.86 ± 1.00	7.67 ± 1.00	71.21 ± 5.00
RIR		1.14	-4.14	0.57	31.36
Audio de voz	250	1.16	-3.81	2.84	29.40
JND		1.14 ± 0.06	-4.14 ± 1.00	0.57 ± 1.00	31.26 ± 5.00
RIR		1.75	-4.59	-1.55	28.11
Audio de voz	500	1.59	-5.38	-0.18	22.50
JND		1.75 ± 0.09	-4.59 ± 1.00	-1.55 ± 1.00	28.11 ± 5.00
RIR		1.45	-2.77	0.39	36.00
Audio de voz	1000	1.69	0.01	2.16	50.10
JND		1.45 ± 0.07	-2.77 ± 1.00	0.39 ± 1.00	36.00 ± 5.00
RIR		1.63	-2.17	0.63	37.38
Audio de voz	2000	1.27	-0.05	3.35	49.70
JND		1.63 ± 0.08	-2.17 ± 1.00	0.63 ± 1.00	37.38 ± 5.00
RIR		1.08	-0.53	2.91	45.60
Audio de voz	4000	1.19	0.82	3.90	55.60
JND		1.08 ± 0.05	-0.53 ± 1.00	2.91 ± 1.00	45.60 ± 5.00
RIR		1.19	0.82	3.90	55.60
Audio de voz	8000	0.55	5.55	9.14	78.20
JND		1.19 ± 0.06	0.82 ± 1.00	3.90 ± 1.00	55.60 ± 5.00

Con el fin de proporcionar una mejor comprensión de las tablas anteriores, se decidió calcular las diferencias entre los valores estimados por ambas redes neuronales y los descriptores obtenidos mediante métodos convencionales. Estos resultados se muestran en la tabla 17.

Tabla 17. Diferencias obtenidas entre el cálculo y la estimación de los descriptores en ambas posiciones para la Sala 1.

Posición	Banda [Hz]	Diferencia entre valor calculado y estimado			
		T_{30} [s]	C_{50} [dB]	C_{80} [dB]	D_{50} [%]
1	125	0.07	0.07	0.31	0.94
2		0.44	1.94	5.74	11.49
1	250	0.14	4.41	1.96	23.76
2		0.02	0.33	2.27	1.96
1	500	0.22	0.21	3.31	2.71
2		0.16	0.79	1.37	5.61
1	1000	0.04	1.50	0.28	6.78
2		0.24	2.78	1.77	14.10
1	2000	0.20	3.91	4.33	20.04
2		0.36	2.12	2.72	12.32
1	4000	0.13	2.57	1.98	13.98
2		0.11	1.35	0.99	10.00
1	8000	0.82	4.37	5.80	23.27
2		0.64	4.73	5.24	22.60

Los valores estimados para esta primera sala fueron muy cercanos a los valores esperados, e incluso muchos de ellos estuvieron dentro del límite de los JND. Ambas posiciones mostraron mediciones consistentes, excepto en el caso del descriptor D_{50}, donde se observaron dispersiones más significativas.

Se pudo observar que las mejores estimaciones se obtuvieron en las bandas cercanas a las frecuencias de la voz. Esto puede atribuirse a la mayor energía presente en estas bandas en comparación con las frecuencias por debajo y por encima del espectro del habla.

5.5.2. Sala 2

Al igual que en la primera sala, se realizaron grabaciones de barridos frecuenciales y audios de voz en dos puntos diferentes de la sala. A continuación, se presentan los resultados obtenidos para cada posición.

En la tabla 18, se muestran los valores estimados por la red neuronal y los valores calculados a partir de las respuestas al impulso para la posición 1.

Por otro lado, la tabla 19 presenta los valores estimados por la red neuronal y los valores obtenidos mediante las respuestas al impulso para la posición 2.

Para una mejor comparación de los resultados, en la tabla 20 se muestran las diferencias entre los valores estimados en ambas posiciones por la red neuronal y los valores

obtenidos a partir de las respuestas al impulso del recinto.

Tabla 18. Valores de los descriptores calculados a partir de una respuesta al impulso contra los estimados a partir de la grabación de voz con la red neuronal para la posición 1.

Obtenido de	Banda [Hz]	T_{30} [s]	C_{50} [dB]	C_{80} [dB]	D_{50} [%]
RIR		1.02	2.60	5.58	60.59
Audio de voz	125	0.95	1.51	3.29	48.50
JND		1.02 ± 0.05	2.60 ± 1.00	5.58 ± 1.00	60.59 ± 5.00
RIR		0.90	-1.75	2.90	40.58
Audio de voz	250	1.39	-0.92	3.07	44.70
JND		0.90 ± 0.05	-1.75 ± 1.00	2.90 ± 1.00	40.58 ± 5.00
RIR		1.54	-3.45	-0.69	33.30
Audio de voz	500	1.61	-2.84	-0.33	34.20
JND		1.54 ± 0.08	-3.45 ± 1.00	-0.69 ± 1.00	33.30 ± 5.00
RIR		1.64	-2.42	0.68	39.76
Audio de voz	1000	1.70	0.63	1.60	53.60
JND		1.64 ± 0.08	-2.42 ± 1.00	0.68 ± 1.00	39.76 ± 5.00
RIR		1.71	-3.35	-0.50	34.26
Audio de voz	2000	1.34	-0.40	3.46	47.70
JND		1.71 ± 0.09	-3.35 ± 1.00	-0.50 ± 1.00	34.26 ± 5.00
RIR		1.15	-0.98	2.24	44.81
Audio de voz	4000	0.91	2.84	5.73	65.80
JND		1.15 ± 0.06	-0.98 ± 1.00	2.24 ± 1.00	44.81 ± 5.00
RIR		1.31	2.96	5.52	66.33
Audio de voz	8000	0.61	7.34	10.20	84.40
JND		1.31 ± 0.07	2.96 ± 1.00	5.52 ± 1.00	66.33 ± 5.00

Tabla 19. Valores de los descriptores calculados a partir de una respuesta al impulso contra los estimados a partir de la grabación de voz con la red neuronal para la posición 2.

Obtenido de	Banda [Hz]	T_{30} [s]	C_{50} [dB]	C_{80} [dB]	D_{50} [%]
RIR	125	1.05	1.12	6.05	56.37
Audio de voz		0.86	4.83	7.13	75.30
JND		1.05 ± 0.05	1.12 ± 1.00	6.05 ± 1.00	56.37 ± 5.00
RIR	250	1.31	0.52	4.70	50.50
Audio de voz		1.24	-3.53	1.91	30.70
JND		1.31 ± 0.07	0.52 ± 1.00	4.70 ± 1.00	50.50 ± 5.00
RIR	500	2.10	-4.50	-1.88	31.38
Audio de voz		1.81	-3.50	-0.78	30.90
JND		2.10 ± 0.11	-4.50 ± 1.00	-1.88 ± 1.00	31.38 ± 5.00
RIR	1000	1.93	-2.37	0.84	39.96
Audio de voz		1.70	-1.83	0.76	39.60
JND		1.93 ± 0.10	-2.37 ± 1.00	0.84 ± 1.00	39.96 ± 5.00
RIR	2000	1.72	-2.80	0.35	38.84
Audio de voz		1.40	-0.66	2.68	46.20
JND		1.72 ± 0.09	-2.80 ± 1.00	0.35 ± 1.00	38.84 ± 5.00
RIR	4000	1.21	-0.43	2.76	47.08
Audio de voz		0.96	0.41	4.15	52.40
JND		1.21 ± 0.06	-0.43 ± 1.00	2.76 ± 1.00	47.08 ± 5.00
RIR	8000	0.81	1.69	5.91	60.17
Audio de voz		0.60	4.71	8.20	74.70
JND		0.81 ± 0.04	1.69 ± 1.00	5.91 ± 1.00	60.17 ± 5.00

Tabla 20. Diferencias obtenidas entre el cálculo y la estimación de los descriptores en ambas posiciones para la Sala 2.

Posición	Banda [Hz]	Diferencia entre valor calculado y estimado			
		T_{30} [s]	C_{50} [dB]	C_{80} [dB]	D_{50} [%]
1	125	0.07	1.09	2.29	12.09
2		0.19	3.71	1.08	18.93
1	250	0.49	0.83	0.17	4.12
2		0.07	4.05	2.79	19.80
1	500	0.07	0.61	0.36	0.90
2		0.29	1.00	1.10	0.48
1	1000	0.06	3.05	0.92	13.84
2		0.23	0.54	0.08	0.36
1	2000	0.37	2.95	3.96	13.44
2		0.32	2.14	2.33	7.36
1	4000	0.24	3.82	3.49	20.99
2		0.25	0.84	1.39	5.32
1	8000	0.70	4.38	4.68	18.07
2		0.21	3.02	2.29	14.53

Continuando con la tendencia observada en la sala anterior, se obtuvieron valores estimados muy similares a los calculados mediante métodos convencionales. Una vez más, se encontraron numerosos resultados dentro del límite de los JND, lo que indica una buena precisión en la estimación. Sin embargo, al igual que en la sala anterior, el descriptor que muestra las mayores variaciones es el D_{50}. Esta consistencia en los resultados refuerza la confianza en el desempeño de la red neuronal en la estimación de los parámetros acústicos en diferentes salas.

5.5.3. Sala 3

En la tercera sala evaluada, se siguieron los mismos procedimientos que en las salas anteriores, grabando audios de voz y obteniendo las respuestas al impulso correspondientes. Los resultados estimados por la red neuronal para la primera posición se presentan en la tabla 21.

Tabla 21. Valores de los descriptores calculados a partir de una respuesta al impulso contra los estimados a partir de la grabación de voz con la red neuronal para la posición 1.

Obtenido de	Banda [Hz]	T_{30} [s]	C_{50} [dB]	C_{80} [dB]	D_{50} [%]
RIR		0.71	-1.30	3.31	44.96
Audio de voz	125	0.78	4.67	5.56	74.50
JND		0.71 ± 0.04	-1.30 $\pm$ 1.00	3.31 $\pm$ 1.00	44.96 $\pm$ 5.00
RIR		0.86	-0.63	3.56	45.65
Audio de voz	250	1.50	2.52	3.05	64.10
JND		0.86 ± 0.04	-0.63 $\pm$ 1.00	3.56 $\pm$ 1.00	45.65 $\pm$ 5.00
RIR		1.69	-4.73	-1.67	27.92
Audio de voz	500	1.43	-3.45	-0.21	31.10
JND		1.69 ± 0.08	-4.73 $\pm$ 1.00	-1.67 $\pm$ 1.00	27.92 $\pm$ 5.00
RIR		1.32	-2.39	0.60	38.03
Audio de voz	1000	1.60	-1.64	1.00	40.70
JND		1.32 ± 0.07	-2.39 $\pm$ 1.00	0.60 $\pm$ 1.00	38.03 $\pm$ 5.00
RIR		1.30	-2.40	0.74	38.84
Audio de voz	2000	1.18	0.39	2.25	52.30
JND		1.30 ± 0.07	-2.40 $\pm$ 1.00	0.74 $\pm$ 1.00	38.84 $\pm$ 5.00
RIR		0.92	-0.23	3.62	49.35
Audio de voz	4000	0.81	2.15	5.15	62.10
JND		0.92 ± 0.05	-0.23 $\pm$ 1.00	3.62 $\pm$ 1.00	49.35 $\pm$ 5.00
RIR		1.24	4.03	6.12	70.59
Audio de voz	8000	0.54	5.89	9.02	79.50
JND		0.54 ± 0.03	4.03 $\pm$ 1.00	6.12 $\pm$ 1.00	70.59 $\pm$ 5.00

Tabla 22. Valores de los descriptores calculados a partir de una respuesta al impulso contra los estimados a partir de la grabación de voz con la red neuronal para la posición 2.

Obtenido de	Banda [Hz]	T_{30} [s]	C_{50} [dB]	C_{80} [dB]	D_{50} [%]
RIR	125	0.91	-0.78	2.99	47.77
Audio de voz		0.62	4.53	5.55	73.90
JND		0.91 ± 0.05	-0.78 ± 1.00	2.99 ± 1.00	47.77 ± 5.00
RIR	250	1.10	-1.84	2.09	40.12
Audio de voz		1.18	-1.81	-1.40	39.70
JND		1.10 ± 0.06	-1.84 ± 1.00	2.09 ± 1.00	40.12 ± 5.00
RIR	500	1.58	-4.07	-0.82	30.99
Audio de voz		1.48	-1.46	2.88	41.70
JND		1.58 ± 0.08	-4.07 ± 1.00	-0.82 ± 1.00	30.99 ± 5.00
RIR	1000	1.49	-2.83	0.19	36.39
Audio de voz		1.54	-2.74	-1.36	34.70
JND		1.54 ± 0.08	-2.83 ± 1.00	0.19 ± 1.00	36.39 ± 5.00
RIR	2000	1.32	-2.48	0.28	36.21
Audio de voz		1.11	-0.29	2.56	48.40
JND		1.32 ± 0.07	-2.48 ± 1.00	0.28 ± 1.00	36.21 ± 5.00
RIR	4000	0.98	-0.13	3.36	48.66
Audio de voz		0.77	1.18	4.59	56.80
JND		0.98 ± 0.05	-0.13 ± 1.00	3.36 ± 1.00	48.66 ± 5.00
RIR	8000	1.13	1.78	4.51	59.67
Audio de voz		0.54	4.65	8.69	74.50
JND		1.13 ± 0.06	1.78 ± 1.00	4.51 ± 1.00	59.67 ± 5.00

Tabla 23. Diferencias obtenidas entre el cálculo y la estimación de los descriptores en ambas posiciones para la Sala 3.

Posición	Banda [Hz]	Diferencia entre valor calculado y estimado			
		T_{30} [s]	C_{50} [dB]	C_{80} [dB]	D_{50} [%]
1	125	0.07	5.97	2.25	29.54
2		0.29	5.31	2.56	26.13
1	250	0.64	3.15	0.51	18.45
2		0.08	0.03	3.49	0.42
1	500	0.26	1.28	1.46	3.18
2		0.10	2.61	3.70	10.71
1	1000	0.28	0.75	0.40	2.67
2		0.05	0.09	1.55	1.69
1	2000	0.12	2.79	1.51	13.46
2		0.21	2.19	2.28	12.19
1	4000	0.11	2.38	1.53	12.75
2		0.21	1.31	1.23	8.14
1	8000	0.70	1.86	2.90	8.91
2		0.59	2.87	4.19	14.83

Asimismo, los valores estimados para la segunda posición se encuentran en la tabla 22.

Para facilitar la comparación de los resultados anteriores, se calculan las diferencias entre los valores estimados en ambas posiciones y sus respectivas referencias. Estas diferencias se presentan en la tabla 23.

Una vez más, se observa una mayor dispersión en el descriptor D_{50} en comparación con los otros parámetros, tanto en la sala actual como en las anteriores. Sin embargo, en esta sala también se obtuvieron resultados muy precisos en ambas posiciones.

Después de analizar todos los recintos y examinar las tablas de diferencias en relación con los JND correspondientes a cada descriptor, se puede concluir que la red neuronal es capaz de estimar con gran precisión los descriptores de las salas. A pesar de las diferencias en las dimensiones de los recintos, no se observaron mejoras en la estimación en ninguna de las salas, lo que indica que el algoritmo funciona de manera consistente en diferentes volúmenes de recintos, siempre y cuando los descriptores se encuentren dentro de los valores utilizados durante el entrenamiento.

No se pudo determinar con certeza por qué el descriptor D_{50} presentó una mayor dispersión que los otros tres parámetros. Sin embargo, se estima que esto puede deberse al hecho de que el cálculo de este descriptor no considera una escala de decibelios.

6. CONCLUSIONES

En esta investigación se desarrolló un algoritmo que utiliza procesamiento de señales de habla y redes neuronales convolucionales para la obtención ciega de parámetros acústicos de una sala. Para entrenar el algoritmo, se utilizó una base de datos de grabaciones de voz reverberadas generadas artificialmente con respuestas al impulso de diferentes recintos. El objetivo era estimar los descriptores a partir de las envolventes temporales de estas señales. Además, se agregó ruido rosa a los archivos de audio para simular diferentes relaciones señal/ruido controladas y entrenar la red en condiciones más realistas.

A diferencia de otros algoritmos existentes en el estado del arte, el enfoque utilizado en este trabajo permite estimar varios descriptores acústicos simultáneamente. Esto marca una diferencia significativa, ya que la mayoría de los métodos existentes entrenan el mismo algoritmo repetidamente para obtener un único descriptor en cada caso.

Dada la dificultad de encontrar una base de datos que represente todos los tipos posibles de recintos, se exploraron técnicas de generación de respuestas al impulso sintéticas y aumentación de respuestas reales para abarcar la mayor cantidad de situaciones de medición en campo posibles.

A lo largo de esta investigación, se demostró que es factible implementar un algoritmo capaz de estimar varios parámetros acústicos simultáneamente, ya que todos estos parámetros se derivan de las curvas de decaimiento de las señales de audio utilizadas. Además, se validaron las técnicas utilizadas para la aumentación de respuestas reales y la generación de respuestas sintéticas, y se observó que su uso mejoró significativamente el proceso de entrenamiento al ampliar y homogeneizar la base de datos.

Al utilizar datos de validación, se encontró que la red neuronal entrenada sin agregar ruido rosa obtuvo los mejores resultados de estimación en todas las bandas de frecuencia. En cuanto a las mediciones de campo, se obtuvieron buenas estimaciones de los parámetros, incluso dentro del rango de las diferencias apenas perceptibles (JND, por sus siglas en inglés). Sin embargo, el descriptor D_{50} presentó una mayor dispersión, lo que indica que aún hay margen para mejorar y ajustar el algoritmo.

En conclusión, se demostró que esta técnica de estimación ciega de parámetros acústicos es una alternativa útil para obtener una idea de los valores de los parámetros de un

recinto. Presenta ventajas como su simplicidad, bajo costo, breve tiempo de medición y la posibilidad de prescindir de equipos complejos. Además, no requiere que el recinto tenga condiciones de bajo ruido, a diferencia de las técnicas normadas. Sin embargo, los resultados también señalan que todavía hay mucho por explorar en este campo. La capacidad de estimar múltiples descriptores simultáneamente abre la posibilidad de crear y mejorar muchos algoritmos existentes, y se espera que este trabajo sirva de inspiración en ese sentido.

7. TRABAJO FUTURO

Los resultados obtenidos en esta investigación son considerados altamente satisfactorios, ya que se pudo demostrar la posibilidad de estimar múltiples parámetros acústicos simultáneamente utilizando una única red neuronal. Sin embargo, existen áreas que podrían mejorarse en el proceso. En primer lugar, las técnicas de aumentación utilizadas para las respuestas al impulso solo permiten variar el tiempo de reverberación o la relación directo-reverberado. Esto genera un control preciso sobre el tiempo de reverberación en la base de datos, pero deja los valores de claridad y definición como aleatorios. Como resultado, la red presenta un mayor porcentaje de aciertos en la estimación del T_{30} y menos en el C_{50}, C_{80} y D_{50}. Sería conveniente implementar técnicas de aumentación que también permitan manipular estos parámetros.

Por otro lado, no se realizó un estudio exhaustivo sobre la arquitectura de la red seleccionada. Por lo tanto, es posible que una variación en dicha arquitectura o la implementación de otro tipo de red completamente diferente pueda mejorar las estimaciones. También es importante considerar el entrenamiento del modelo utilizando una base de datos que combine tanto señales con ruido como señales sin ruido. Esto sería útil para evaluar si esta combinación genera mejores estimaciones en escenarios de mediciones reales.

Las mediciones de campo realizadas se llevaron a cabo en condiciones favorables según los estándares establecidos. Es decir, en un entorno con poco ruido presente. En cierto sentido, esto puede haber influido en que la red neuronal entrenada sin agregar ruido a los datos obtuviera mejores resultados en la estimación, como se observó en el análisis de la investigación. Por lo tanto, sería recomendable agregar fuentes externas y controladas de ruido en futuras mediciones para determinar si este enfoque es realmente superior o si el hecho de utilizar ruido rosa durante el entrenamiento permite lograr mejores estimaciones en condiciones adversas.

Por último, es necesario analizar técnicas de aumentación que puedan simular recintos con geometrías diferentes a las paralelepipédicas, y estudiar si el modelo sigue funcionando correctamente en esos casos o si se requiere una optimización específica para dichas situaciones.