UNTREF UNIVERSIDAD NACIONAL DE TRES DE FEBRERO

INGENIERÍA DE SONIDO

Estimación ciega de parámetros acústicos de un recinto†

Autor: Maximiliano Adriel Ortiz
Tutor: Ing. Martin Bernardo Meza

(†) **Tesis para optar por el título de ingeniero de Sonido**.

Febrero 2023

ÍNDICE DE CONTENIDOS

1 INTRODUCCIÓN **1**

1.1 FUNDAMENTACIÓN 1

1.2 OBJETIVOS 2

1.2.1 OBJETIVO GENERAL 2

1.2.2 OBJETIVOS ESPECÍFICOS 2

1.3 ESTRUCTURA DE LA INVESTIGACIÓN 3

2 ESTADO DEL ARTE **4**

2.1 MODELOS DE ESTIMACIÓN CIEGA DE PARÁMETROS ACÚSTICOS 4

3 MARCO TEÓRICO **8**

3.1 RESPUESTA AL IMPULSO DE UNA SALA: RIR 8

3.2 DESCRIPTORES ACÚSTICOS DE UNA SALA 9

3.2.1 TIEMPO DE REVERBERACIÓN: EDT, T10, T20, T30 Y T60 10

3.2.2 CLARIDAD: C_{80} Y C_{50} 11

3.2.3 DEFINICIÓN: D_{50} 12

3.3 RESPUESTAS AL IMPULSO SINTÉTICAS 12

3.4 RELACIÓN DIRECTO-REVERBERADO: DRR 14

3.5 ENVOLVENTE TEMPORAL DE AMPLITUD: TAE 14

3.6 REDES NEURONALES ARTIFICIALES 16

3.6.1 LA NEURONA ARTIFICIAL O PERCEPTRÓN 16

3.6.2 PERCEPTRÓN MULTICAPA 17

3.6.3 ENTRENAMIENTO DE UNA RED NEURONAL 19

3.6.4 REDES NEURONALES CONVOLUCIONALES (CNN) 20

4 METODOLOGÍA **23**

4.1 GENERACIÓN DE BASE DE DATOS 23

4.2 AUMENTACIÓN DE RESPUESTAS AL IMPULSO 23

4.2.1 AUMENTACIÓN DE LA RELACIÓN DIRECTO-REVERBERADO 24

4.2.2 AUMENTACIÓN DEL TIEMPO DE REVERBERACIÓN 25

4.3 CURVA DE DECAIMIENTO DE UNA SEÑAL DE VOZ 27

4.4 OBTENCIÓN DE RESPUESTAS AL IMPULSO 28
4.4.1 RESPUESTAS AL IMPULSO SINTÉTICAS 28
4.4.2 RESPUESTAS AL IMPULSO REALES 29
4.4.3 RESPUESTAS AL IMPULSO AUMENTADAS 31
4.5 BASE DE DATOS DE SEÑALES DE VOZ 32
4.6 CÁLCULO DE LOS DESCRIPTORES DE LAS RESPUESTAS AL IMPULSO 33
4.7 OBTENCIÓN DE LAS ENVOLVENTES TEMPORALES DE AMPLITUD: TAE 33
4.7.1 GENERACIÓN DE TAE SIN RUIDO 33
4.7.2 GENERACIÓN DE TAE CON RUIDO ROSA 34
4.8 MODELO PROPUESTO 35
4.9 EVALUACIÓN DEL MODELO 37
4.10 COMPARACIÓN DEL MODELO CON MEDICIONES DE CAMPO 38

5 RESULTADOS Y ANÁLISIS **41**
5.1 ANÁLISIS DE LA BASE DE DATOS DE RESPUESTAS AL IMPULSO 41
5.2 ENTRENAMIENTO DE LOS MODELOS 47
5.3 EVALUACIÓN DEL MODELO CON AUDIOS DESCONOCIDOS 48
5.3.1 ANÁLISIS T_{30} 49
5.3.2 ANÁLISIS C_{50} 50
5.3.3 ANÁLISIS C_{80} 50
5.3.4 ANÁLISIS D_{50} 51
5.4 EVALUACIÓN DEL MODELO CON MEDICIONES REALES 52
5.4.1 SALA 1 52
5.4.2 SALA 2 55
5.4.3 SALA 3 58

6 CONCLUSIONES **62**

7 TRABAJO FUTURO **64**

BIBLIOGRAFÍA **65**

ÍNDICE DE FIGURAS

Figura 1 Ejemplo de respuesta al impulso de una sala. 9
Figura 2 Decaimiento de una respuesta al impulso y sus aproximaciones por cuadrados mínimos para los descriptores EDT, T10, T20 y T30. 11
Figura 3 Pasos para la obtención de la TAE. 15
Figura 4 Esquema de una neurona artificial. 16
Figura 5 Funciones de activación. 17
Figura 6 Perceptrón multicapa. 18
Figura 7 Esquema de un filtro de convolución. 22
Figura 8 Proceso de aumentación del DRR. 25
Figura 9 Diferencias en la curva de decaimiento para un audio limpio contra uno reverberado. 28
Figura 10 Respuestas al impulso con distintos tiempos de reverberación generadas de forma sintética. 29
Figura 11 Esquema de medición de respuestas al impulso del recinto Great Hall. 30
Figura 12 Esquema de medición de respuestas al impulso del recinto Octagon. 30
Figura 13 Esquema de medición de respuestas al impulso del recinto Classroom. 31
Figura 14 Banco de filtros para el análisis de aumentación de tiempo de reverberación de las respuestas al impulso. 32
Figura 15 Diagrama en bloques del modelo propuesto. 36
Figura 16 Imagen de una de las salas medidas. 39
Figura 17 Esquema de medición de la sala 1. 39
Figura 18 Esquema de medición de la sala 2. 40
Figura 19 Esquema de medición de la sala 3. 40
Figura 20 Gráfico de dispersión del $T30_{mid}$ contra el DRR_{mid} para las RIR sintéticas. 42
Figura 21 Gráfico de dispersión del $T30mid$ contra el DRR_{mid} para las RIR de salas reales. 43
Figura 22 Gráfico de dispersión del $T30mid$ contra el DRR_{mid} para las RIR aumentadas. 44

Figura 23 Gráfico de dispersión del $T30mid$ contra el DRR_{mid} para toda la base de datos. 45
Figura 24 Boxplot de los valores de T_{30} por bandas de octava de toda la base de datos. 45
Figura 25 Boxplot de los valores de DRR por bandas de octava de toda la base de datos. 46
Figura 26 Curvas de entrenamiento para la banda de 1000 Hz de los modelos sin ruido y con ruido respectivamente. 47

ÍNDICE DE TABLAS

Tabla 1 Tiempos de reverberación por bandas de frecuencia para cada sala. 30

Tabla 2 Arquitectura de red propuesta. 37

Tabla 3 Valores máximos y mínimos de los descriptores T_{30} y DRR mid. . . 44

Tabla 4 Cantidad de RIRs en la base de datos y total de horas de entrenamiento por bandas de frecuencia. 46

Tabla 5 Valores de Loss y Validation loss obtenidos durante el entrenamiento de las redes neuronales usando la base de datos con ruido y sin ruido 48

Tabla 6 Coeficiente de correlación de Spearman entre las predicciones y el valor real de T_{30} por cada banda y por base de datos utilizada. 49

Tabla 7 Coeficiente de correlación de Spearman entre las predicciones y el valor real de C_{50} por cada banda y por base de datos utilizada. 50

Tabla 8 Coeficiente de correlación de Spearman entre las predicciones y el valor real de C_{80} por cada banda y por base de datos utilizada. 51

Tabla 9 Coeficiente de correlación de Spearman entre las predicciones y el valor real de D_{50} por cada banda y por base de datos utilizada. 51

Tabla 10 JND de los descriptores acústicos. 52

Tabla 11 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. 53

Tabla 12 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. 54

Tabla 13 Diferencias obtenidas entre el cálculo y la estimación de los descriptores en ambas posiciones para la Sala 1. 55

Tabla 14 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. 56

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 2. 57
Tabla 16 Diferencias obtenidas entre el cálculo y la estimación de los descriptores en ambas posiciones para la Sala 2. 58
Tabla 17 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. 59
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 2. 60
Tabla 19 Diferencias obtenidas entre el cálculo y la estimación de los descriptores en ambas posiciones para la Sala 3. 61

Resumen

El objetivo de esta investigación es estudiar un método que permita la obtención de parámetros acústicos a partir de señales de voz grabadas en un recinto, evitando la medición de respuestas al impulso de la sala. Para esto, se propone un modelo, inspirado en el estado de arte actual, que utiliza redes neuronales convolucionales para dicho fin. A diferencia de los anteriormente planteados, el de esta tesis tiene la particularidad de poder obtener múltiples parámetros a la vez. Para entrenar esta red, se ingresa con la información de la envolvente temporal de amplitud (TAE) de una señal de habla reverberada, de la cual se obtiene a su salida 4 descriptores correspondientes a los parámetros buscados. En concreto, en esta investigación se buscan estimar los descriptores: tiempo de reverberación (T_{30}), claridad (C_{50} y C_{80}) y definición (D_{50}). Dada la escasa cantidad de datos disponibles, para lograr entrenar el sistema propuesto se genera una base de datos de audios de voz reverberados a partir de la convolución de señales de voz anecoicas con un banco de respuestas al impulso reales, otros generados artificialmente y otros que salen a partir de la aumentación de datos de los impulsos reales. Finalmente, para probar la viabilidad del modelo se comparan con los descriptores obtenidos en una sala mediante métodos convencionales en contraste con el propuesto. Se realizó un estudio de la correlación entre las diferencia de los resultados esperados y los valores obtenidos con el conjunto de datos de pruebas y se obtuvo que la misma es de más del 90 % en cada banda de frecuencia, excepto la de 125 Hz.

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

Abstract

The aim of this research is to study a method that allows obtaining acoustic parameters from speech signals recorded in an room, avoiding the measurement of impulse responses of the room. For this purpose, inspired by the current state of the art, a model which uses convolutional neural networks for this task is proposed. Unlike the ones previously proposed, the model of this thesis has the particularity of being able to obtain multiple parameters at the same time. To train this network, the information of the temporal amplitude envelope (TAE) of a reverberated speech signal is entered, from which 4 descriptors corresponding to the searched parameters are obtained at its output. Specifically, in this research we seek to estimate the descriptors: reverberation time (T_{30}), clarity (C_{50} and C_{80}) and definition (D_{50}). Given the limited amount of available data, in order to train the proposed system, a database of reverberated speech audios is generated from the convolution of anechoic speech signals with a bank of real impulse responses, others generated artificially and others that come out from the data augmentation of the real impulses. Finally, to test the feasibility of the model, they are compared with the descriptors obtained in a room by conventional methods in contrast to the proposed one. A study of the correlation between the difference between the expected results and the values obtained with the test data set was carried out and it was obtained that the correlation is more than 90 % in every frequency band, except 125 Hz.

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

1. INTRODUCCIÓN

1.1. FUNDAMENTACIÓN

Para todo aquel que le interese realizar el estudio acústico de un recinto, la medición de parámetros acústicos de una sala es una herramienta fundamental si se busca entender objetivamente el comportamiento de la misma y poder hacer todos los ajustes necesarios para conseguir la calidad y tipo de sonidos deseados. Ya sea que se busque un tiempo de reverberación bajo y una buena inteligibilidad dentro de una sala de reuniones o un recinto que permita escuchar perfectamente las señales sonoras de emergencia [1][2], los descriptores acústicos son la mejor opción [3].

Las técnicas de obtención de los parámetros objetivos son bien conocidas y están normadas, ya sea para el cálculo del índice de transmisión de voz (STI), el tiempo de reverberación (TR), claridad (C_{80}), entre otras [4] [5]. Los descriptores anteriormente mencionados (y la mayoría de los demás no citados), se calculan a partir de conocer la respuesta al impulso de una sala. Esto es así porque, por teoría de señales y sistemas, se sabe que a partir de la respuesta al impulso de un sistema lineal e invariante en el tiempo (LTI) se puede obtener el comportamiento del mismo frente a un estímulo conocido [6].

Existen diversos métodos, unos más sofisticados que otros, para calcular la respuesta al impulso de una sala. El procedimiento de obtención consiste en generar una señal de corta duración y gran amplitud, que sea capaz de excitar la sala en todas las frecuencias. Entre los métodos más conocidos están: disparo en blanco, explosión de globos, aplausos, choque con maderas, barrido frecuencial, entre otros [7]. Sea cual sea el método que se utilice, todos estos se ven afectados por los ruidos que pueda haber en la sala, ya sea el provocado por gente en su interior o ruidos de fondo de maquinarias, tráfico o demás fuentes. Esto es grave en situaciones donde se desea medir la respuesta de la sala pero no es posible vaciarla, como, por ejemplo, en la estación de un subterráneo. Por otra parte, muchos de estos métodos son limitados en cuanto al espectro de frecuencias que pueden excitar (como, por ejemplo, la explosión de globos en el cual el diámetro del mismo determina la frecuencia más baja en la cual puede generar energía) y, además, la mayoría de ellos no son replicables, lo cual genera diferentes resultados al repetir el mismo procedimiento [8].

Por esta razón, en esta investigación se propone un método de medición de pará-

metros acústicos que consiste en la estimación de los mismos a partir de una señal de voz grabada con la reverberación propia del recinto en cuestión. Este tipo de medición se conoce en el estado del arte como estimación ciega de parámetros acústicos porque, a diferencia de los métodos convencionales, no se utilizan equipos para medir la respuesta al impulso de la sala sino que se modela a partir de un audio de voz reverberado y algoritmos con redes neuronales [9].

Para ello, la investigación se aborda desde el enfoque de la ingeniería de audio, estudiando y entendiendo las limitaciones de los modelos actuales, planteando mejoras y un estudio detallado de los resultados obtenidos.

1.2. OBJETIVOS

1.2.1. OBJETIVO GENERAL

El objetivo general de esta investigación es implementar un algoritmo de redes neuronales convolucionales que permita estimar los parámetros acústicos de una sala a partir de una grabación de señal de voz reverberada.

1.2.2. OBJETIVOS ESPECÍFICOS

Entre los objetivos específicos, se pueden mencionar:

- Revisar las distintas técnicas utilizadas para la obtención de parámetros acústicos cuando no se cuenta con la respuesta al impulso de la sala.
- Diseñar e implementar un modelo que permita la obtención de parámetros acústicos de forma ciega utilizando redes neuronales en el lenguaje de programación Python.
- Generación de una base de datos de voces reverberadas a partir de un banco de impulsos reales y generados artificialmente.
- Realizar técnicas de aumentación de datos para las respuestas al impulso reales.
- Optimizar el sistema propuesto y comparar los resultados obtenidos con los calculados con los métodos convencionales.
- Realizar una medición empírica de una sala y contrastar los resultados obtenidos.

1.3. ESTRUCTURA DE LA INVESTIGACIÓN

En el capítulo 2 se presenta el estado del arte en el campo de la estimación ciega de parámetros acústicos. En el capítulo 3 se detalla el marco teórico necesario para el seguimiento y comprensión de este trabajo. En este se abordan las siguientes temáticas: la respuesta al impulso de una sala (RIR), síntesis de respuestas al impulso, técnicas de aumentación de datos para respuestas al impulso, los parámetros acústicos que se van a estimar (a saber: T_{30}, C_{50}, C_{80} y D_{50}), el concepto de envolvente temporal de amplitud (TAE) y, por último, la aplicación de redes neuronales convolucionales y algoritmos de aprendizaje junto con las principales técnicas de procesamiento de las señales para su entrenamiento. En el capítulo 4 se especifica la metodología seguida a lo largo de este trabajo, y se brinda toda la información necesaria para replicar los experimentos realizados. En el capítulo 5 se presentan los resultados de los experimentos y se hace un análisis crítico de los mismos. En el capítulo 6 se exponen las conclusiones generales del trabajo y, por último, en el capítulo 7 se proponen líneas futuras de investigación relacionadas con la presente investigación.

2. ESTADO DEL ARTE

2.1. MODELOS DE ESTIMACIÓN CIEGA DE PARÁMETROS ACÚSTICOS

Hoy en día, los métodos de obtención de los parámetros acústicos para caracterizar una sala son bien conocidos y se detallan en el marco de diversas normas. Entre los más famosos se encuentran: EDT, T_{10}, T_{20}, T_{30}, C_{50}, C_{80}, D_{50}, Ts [4], STI [5], entre otros.

Todos los parámetros anteriormente nombrados (y la mayoría de los no citados también), tienen la particularidad de que se calculan a partir de la respuesta al impulso de la sala (RIR) a caracterizar. Si bien con el tiempo las técnicas para la obtención de las RIRs se han ido perfeccionando, todavía representan un costo muy grande. Esto se debe, principalmente, en la dificultad de conseguir los equipos para su medición y la complejidad de cumplir las condiciones que exige la normativa. La obtención de respuestas al impulso se ven fuertemente afectadas por la presencia de ruido de fondo en la sala; siendo muchas veces imposible conseguir los 45 dB de rango dinámico entre la señal de medición y el piso de ruido que aconseja la norma ISO 3382 [4].

Conociendo estas limitaciones para la obtención de una RIR, muchos autores comenzaron a buscar métodos para calcular los parámetros acústicos que no dependan de esta. El primer paso que se dio en este campo fue en el año 2007 en una investigación presentada por Kendrick et. al. [10], en la cual los autores presentaron un método para estimar el tiempo de reverberación de una sala a partir de un audio grabado en el lugar. Al no obtenerse a partir de la RIR, los autores denominaron este método como una estimación ciega. Para la norma ISO 3382 [4], este descriptor se encuentra a partir de la pendiente que se obtiene en la curva de decaimiento de una respuesta al impulso al aproximar mediante cuadrados mínimos una cierta cantidad de disminución de decibeles (de -5 a -15 para el descriptor T_{10}, -5 a -25 para el T_{20} y -5 a -35 para el T_{30}). Tomando esta lógica, los autores descubrieron que existe una correlación en el decaimiento de una RIR y en el final de una palabra en un audio de voz o en un acorde tocado en una sala. A partir de esto, grabaron audios reverberados (es decir, capturados en la sala) y separaron los fragmentos en los que habían ciertas curvas de decaimiento similares a las de una RIR. Luego de encontrar estos fragmentos, realizaron una aproximación de este a una pendiente perfecta mediante un estimador de máxima verosimilitud (MLE) y pudieron calcular el tiempo de reverberación de banda completa de la

sala con un buen grado de exactitud.

El método anteriormente descripto es bueno pero tiene tres puntos débiles: no es capaz de estimar el tiempo de reverberación por bandas de frecuencia, sigue siendo afectado por el ruido de fondo y no sirve para determinar otros parámetros acústicos. Teniendo estas consideraciones, los mismos autores presentaron al año siguiente un método que es capaz de calcular los parámetros T_{20}, EDT, C_{80} y Ts por bandas de frecuencia [11]. De mano con el auge de la época por las redes neuronales artificiales, en esta ocasión la estimación se hace a partir de entrenar una red con la envolvente de audios de voz grabados en una sala y filtrados en una cierta banda de frecuencia de preferencia. La salida de esta red (es decir, el valor al que trata de aproximar) es el descriptor de la sala en la banda de frecuencia a la cual se filtró previamente el audio.

Para este punto, la estimación ciega de cualquier parámetro acústico ya era posible pero el método seguía siendo fuertemente dependiente del ruido de fondo en las salas, además de que las redes no eran muy eficientes ya que estimaban un solo parámetro a la vez, siendo necesario entrenar una red para cada parámetro en cuestión. Para tratar de solucionar esto, la *IEEE Audio and Acoustic Signal Processing Technical Committee* creo un desafío llamado *Acoustic Characterisation of Environments Challenge* (ACE Challenge) [12] en el año 2015. El objetivo de este desafío era evaluar algoritmos de última generación para la estimación ciega del tiempo de reverberación (RT) y la relación directo-reverberado (DRR) de una sala a partir de un audio de voz y promover el área emergente de investigación en este campo. En este desafío se proveía a los participantes de una base de datos de respuestas al impulso grabadas en 5 salas y audios de voz anecoicos para poder reverberarlos [13].

A raíz de este desafío, se presentaron diversos modelos novedosos para el cálculo de estos dos parámetros. Entre ellos, se destacaron los propuestos por Parada et. al. [14] que utilizaron redes neuronales recurrentes para ir estimando tanto el RT como el DRR en pequeños sectores del mismo audio aprovechando la propiedad de memoria de este tipo de redes, el de Prego et. al [15] que calcula los parámetros a partir del decaimiento de la señal de voz filtrada en muchas bandas de frecuencia y promediando todos los valores de RT y DRR encontrados y, por último, el propuesto por Loellman et. al. [16] que utiliza un MLE para estimar los valores de RT y DRR. Estos dos últimos métodos se podrían considerar como una expansión del propuesto por Kendrick et. al. [10]

Si bien todos estos nuevos modelos enriquecieron el conocimiento dentro de esta área, seguían fallando frente a audios con una relación señal a ruido muy baja y al probarlos con grabaciones provenientes de salas con características diferentes a las que se utilizaron para el entrenamiento. Frente a esto, se hace evidente la falta de una base de datos con la suficiente cantidad de respuestas al impulso como para poder cubrir la mayor cantidad de casos posibles durante el entrenamiento de la red neuronal. Como la obtención de tantas RIRs resulta muy costoso, Bryan propuso un método de aumentación de respuestas al impuslo con la finalidad de obtener la mayor cantidad de RIRs posibles y así poder modelar muchas salas con el menor esfuerzo posibles [17]. La utilización de esta nueva base de datos generó una amplia mejora en los modelos previamente propuestos por otros autores, tanto para audios con mucho ruido como para los audios tomados en salas distintas a las del entrenamiento.

Hasta este punto, todos los modelos presentados entrenaban una red o realizaban algún tratamiento a la señal para calcular únicamente un parámetro acústico a la vez (posiblemente inspirados por el primer método de todos). En el caso de los métodos con redes neuronales, se utilizaba toda una red con una estimación para el valor del RT y otra red distinta cuyo resultado era el DRR. No fue hasta el 2021 que Duangpummet et. al. [9] propusieron un método para estimar varios parámetros acústicos a la vez. El modelo se entrena con la envolvente de una señal de audio grabada en la sala y estima el RT por bandas de frecuencias y, a partir de ese valor, sintetiza una respuesta al impulso suponiendo que la misma se puede representar como una exponencial decreciente con cierto decaimiento que está dado por el tiempo de reverberación. A partir de la obtención de esta RIR sintética, es capaz de calcular todos los descriptores con los métodos convencionales de la ISO 3382.

Si bien este último modelo es revolucionario por tener la capacidad de estimar varios descriptores a la vez, asume una hipótesis que no es cierta en todos los casos. El principal problema es que esta red solamente estima un valor de tiempo de reverberación y con este reconstruye una respuesta al impulso con una pendiente ideal, de la cual extrae los demás parámetros. Al hacer esto, se comete el error de suponer que todos los descriptores tienen una correlación perfecta con el tiempo de reverberación, lo cual en muchos casos puede no ser cierto. Esto es así porque existen infinitas respuestas al impulso que responden a un mismo tiempo de reverberación; por lo cual, no es correcto extrapolar directamente todos

los demás parámetros de una respuesta sintética. Para que esto funcione, la RIR reconstruida debería brindar más información de la sala de forma tal que su envolvente se parezca más a la que se podría haber obtenido en una medición real dentro de ese recinto.

3. MARCO TEÓRICO

3.1. RESPUESTA AL IMPULSO DE UNA SALA: RIR

Por teoría de señales y sistemas, se sabe que todo sistema lineal e invariante en el tiempo (LTI) puede describirse a través de su respuesta al impulso. Esta característica es muy útil para muchas áreas dentro de la ingeniería y la acústica no es la excepción.

Si definimos un conjunto conformado por una sala que en su interior cuenta con un micrófono y una fuente, es posible encontrar la respuesta al impulso *h(n)* para poder caracterizar el recinto y, con esta, calcular ciertos parámetros que nos permitan entender las características acústicas del mismo. Para su obtención, es necesario excitar al sistema con un impulso infinitamente angosto (delta de Dirac), lo cual nos dará diferentes valores para cada posición fuente-receptor que se utilice dentro de la sala.

Es importante aclarar que este sistema se tiene que asumir como LTI para poder utilizar el modelo de obtención de respuesta al impulso, pero que no siempre cumple con estas características. No obstante, esta asunción no afecta en gran medida al cálculo de los descriptores.

En este caso, como la captación del sistema se hace a partir de un micrófono de medición, la excitación del mismo debe ser acústica. Los métodos para la obtención de esta respuesta son variados, pero entre ellos se destacan:

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

De todos los métodos anteriormente mencionados, el del LSS es el más popularizado actualmente ya que, a diferencia de los demás, permite un mayor control y repetibilidad de la medición al poder manipular las características del estímulo como: duración, rango de frecuencias y amplitud. Otra de sus ventajas es que se pueden tomar múltiples mediciones

y promediarlas para mejorar la relación señal-ruido de la respuesta al impulso y, por otro lado, también permite obtener la distorsión que tiene el sistema de medición utilizado [18].

A modo ilustrativo, en la Figura 1 se puede observar la respuesta al impulso de una sala generada de forma sintética (este método de generación artificial de RIRs se definirá más adelante en este mismo capítulo).

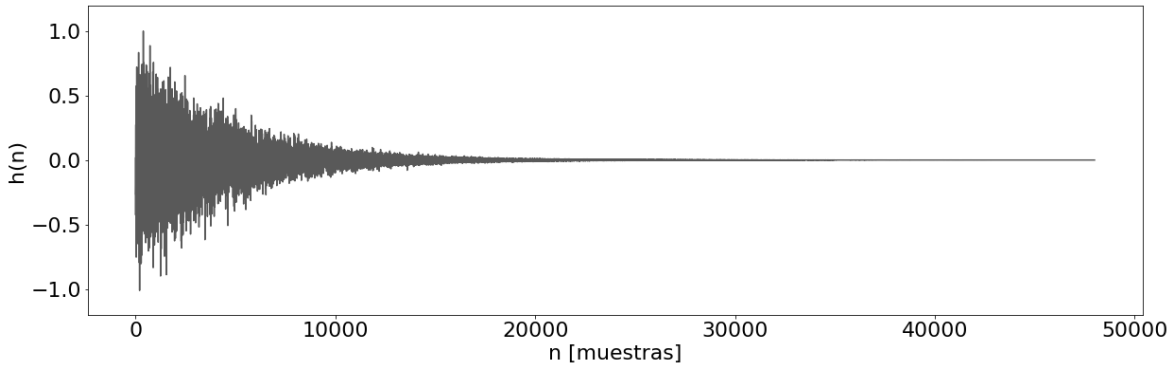

Figura 1. Ejemplo de respuesta al impulso de una sala.

Entendiendo a la sala como un sistema LTI y conociendo su respuesta al impulso, es posible obtener la señal que captaría un micrófono frente a un estimulo generado en su interior, en el punto concreto donde se obtuvo la respuesta. En concreto, esto equivale a la convolución entre la respuesta al impulso de la sala y la señal emitida por la fuente, tal como se observa en la ecuación 1.

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

Donde *h(t)* es la RIR, *s(t)* es el estímulo o señal de entrada e *y(t)* es la respuesta del sistema, a la cual llamaremos, de ahora en más, señal reverberada.

Esto quiere decir que, si se cuenta con un estímulo anecoico, es posible convolucionarlo con una RIR y obtener como respuesta una señal embebida con todas las características acústicas del recinto; como si fuera grabada dentro del mismo.

3.2. DESCRIPTORES ACÚSTICOS DE UNA SALA

En el siguiente apartado se definen los descriptores utilizados en la presente investigación, los mismos se toman a partir de las deficiones dadas por la norma ISO 3382 [4] para el tiempo de reverberación $(T30)$, claridad (C_{50}, C_{80}) y definición (D_{50}).

3.2.1. TIEMPO DE REVERBERACIÓN: EDT, T10, T20, T30 Y T60

El tiempo de reverberación es uno de los descriptores más conocidos e importantes de un recinto. El mismo se define como la duración de tiempo requerida para que la densidad de energía sonora, promediada en el espacio de la sala, disminuya en 60 dB luego de que la emisión de la fuente haya cesado. A este descriptor se lo conoce como T_{60}.

En prácticamente todas las situaciones, las condiciones de ruido de la sala no permiten que la fuente sea capaz de emitir señal que tenga un rango dinámico de más de 60 dB desde el máximo al piso de ruido para el cálculo del descriptor. Por esta razón, surge la necesidad de hacer el cálculo en rangos más acotados y proyectarlo a la disminución deseada.

Por tanto, dada una RIR conocida, el descriptor T_{60} se deriva del momento en que su curva de decaimiento suavizada cumple con las siguientes condiciones:

- EDT: El rango de la curva desde el máximo de la señal hasta 10 dB por debajo.
- T_{10}: El rango de la curva desde 5 dB y 15 dB por debajo del máximo.
- T_{20}: El rango de la curva desde 5 dB y 25 dB por debajo del máximo.
- T_{30}: El rango de la curva desde 5 dB y 35 dB por debajo del máximo.

La curva de decaimiento de una RIR se obtiene, simplemente, a partir del cuadrado de la misma pasado a escala de decibelios. No obstante, los descriptores se calculan a partir del suavizado de este decaimiento. Existen diversos métodos para suavizar la señal, entre los más conocidos se encuentran el filtrado de media móvil [19], la transformada de Hilbert [20] y el método de integración de Schroeder [21].

En esta investigación se utiliza el método de integración de Schroeder. Dada una RIR conocida llamada h(t), es posible obtener su versión suavizada $h_s(t)$ (también llamada envolvente) a partir de la ecuación 2.

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

Es importante aclarar que el método no contempla la presencia de ruido en la respuesta al impulso, por esta razón toma su límite de integración hasta infinito (lo cual se traduciría al largo total del audio en un ejemplo concreto). En la realidad esto nunca será así ya que el

ruido es intrínseco y variable en toda medición. Por esto es necesario tomar un criterio de recorte de la curva de decaimiento. En concreto, en esta investigación, se utiliza el método de Lundeby [22] para determinar el piso de ruido de la señal y, con eso, obtener un límite superior de integración para usar la fórmula de Schroeder.

Una vez determinada la curva de decaimiento suavizada, los descriptores EDT, T_{10}, T_{20} y T_{30} se pueden calcular a partir de la pendiente de la recta aproximada por cuadrados mínimos usando el rango correspondiente que se describió anteriormente.

Por tanto, siendo m la pendiente de la recta de aproximación por cuadrados mínimos, los descriptores de tiempo de reverberación se determinan a partir de la ecuación 3.

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

Donde m es la pendiente de la recta aproximada por cuadrados mínimos y T_x representa los descriptores EDT, T_{10}, T_{20} o T_{30} según el rango utilizado para estimar la recta.

A modo ilustrativo, en la Figura 2 se puede observar la curva de decaimiento de una RIR y las pendientes aproximadas mediante cuadrados mínimos para cada descriptor.

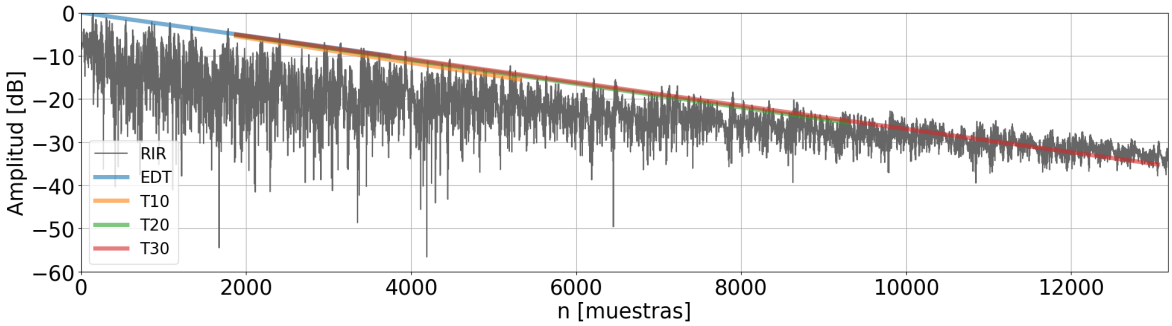

Figura 2. Decaimiento de una respuesta al impulso y sus aproximaciones por cuadrados mínimos para los descriptores EDT, T10, T20 y T30.

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

La claridad es un descriptor que se utiliza para entender la relación de energía directa y tardía que hay en una sala, es decir, da una noción sobre el campo antes y después del arribo de las reflexiones al micrófono. En concreto, el 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 puede calcular mediante la división entre la energía en sus primeros 80 ms contra el resto de la señal,

tal como se observa en la ecuación 4.

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

En paralelo a este descriptor, existe otro llamado C_{50} que se utiliza para determinar la transparencia del habla dentro de la sala. Su cálculo es similar al del C_{80} salvo que cambian sus límites de integración, pasando de 80 a 50 ms.

3.2.3. DEFINICIÓN: D_{50}

La definición, al igual que la claridad, es otro descriptor que se utiliza para determinar la respuesta de la sala comprando la energía en la señal temprana contra la tardía.

En concreto, el D_{50} se utiliza para evaluar la inteligibilidad del habla de las salas de conferencias o aulas, representado en 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 RIR estocástica, es posible utilizar el método de Schroeder [23]. Este método supone que la señal puede ser representada por un ruido blanco gausiano cuya amplitud está modulada por una exponencial decreciente.

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

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

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

Luego, conociendo las RIRs por bandas de frecuencia, la señal original puede ser representada como la suma de todas las bandas. Esto se observa en la ecuación 7.

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

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

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 los de orden inferior (dado 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 [10].

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 ($c_h(t)$), tal 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. Esto se puede observar en la ecuación 9.

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

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

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

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

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 [24]. 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 las muestras correspondientes a la señal directa. Por tanto, todas las muestras que continúan luego de n_d corresponden al campo reverberado por causa de las reflexiones de la onda sonora en el recinto.

3.5. ENVOLVENTE TEMPORAL DE AMPLITUD: TAE

Según lo expuesto en la sección 4.3, se observó que el decaimiento de una señal de voz captada en un recinto guarda información del mismo.

Siguiendo la lógica del cálculo de los descriptores acústicos a través de una RIR, sabemos que la importancia está en la envolvente de la señal y no en su contenido crudo. Por esto, resulta útil generar una versión simplificada de la señal de voz que solo contenga datos de la envolvente del audio. A esta envolvente la vamos a denominar envolvente temporal de amplitud o TAE.

Al estudiar la TAE de la señal de voz en lugar de su versión sin procesar se logran ciertas ventajas: en principio, podemos independizarnos del contenido del audio, ya que lo único importante en el análisis van a ser las partes de decaimiento de la señal de las que se pueden obtener los parámetros acústicos. En segundo lugar, la información se comprime drásticamente, lo cual se traduce en una disminución en la complejidad del entrenamiento de la red neuronal.

Para obtener la TAE se deben realizar los siguientes pasos:

1. **Obtener una señal reverberada**

2. **Filtrar la señal con un filtro pasa banda**

3. **Obtener la transformada de Hilbert de la señal:** esta devuelve la envolvente de una señal.

4. **Filtrar la señal con un filtro pasa bajos con frecuencia de corte en 20 Hz:** esto se hace para reducir la información ya que no nos interesa el contenido de la voz, simplemente la envolvente de la señal.

5. **Resamplear la señal a una frecuencia de muestreo de 40 Hz:** similar a lo anterior, como solo tenemos contenido hasta 20 Hz no es necesario conservar la frecuencia de muestre de 16 kHz de antes. A su vez, esto reduce considerablemente la cantidad de muestras de la señal, lo cual va a ser útil para entrenar más rápido la red neuronal ya que se cuentan con menos datos de entrada.

6. **Normalizar la señal obtenida.**

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

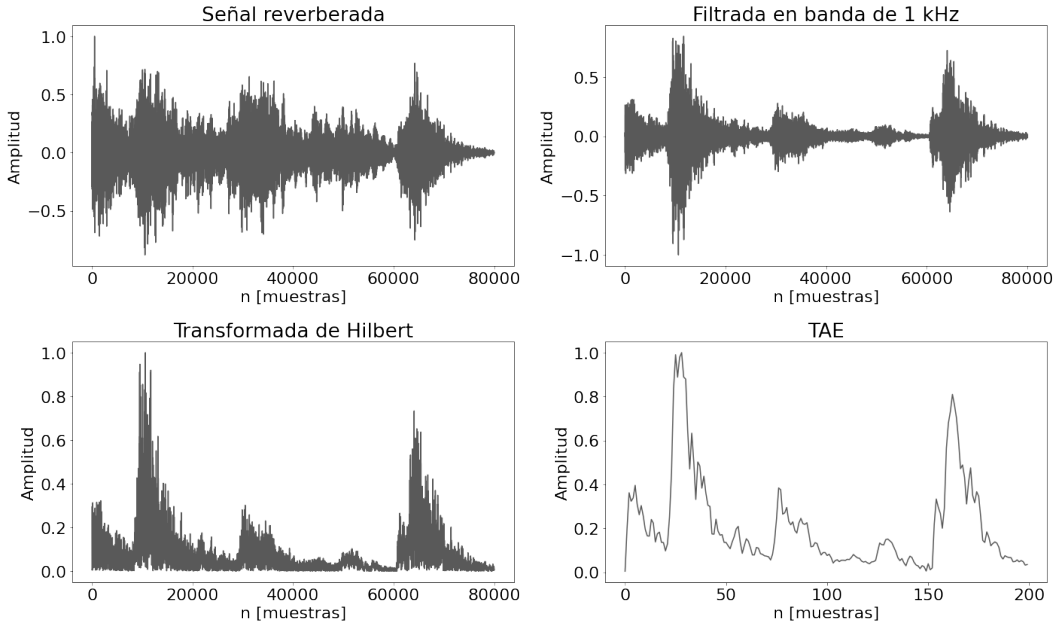

Figura 3. Pasos para la obtención de la TAE.

3.6. REDES NEURONALES ARTIFICIALES

3.6.1. LA NEURONA ARTIFICIAL O PERCEPTRÓN

Las redes neuronales son uno de los muchos algoritmos que componen la rama de investigación del aprendizaje automático (mejor conocido por su nombre en inglés: machine learning) [25].

Para entender el funcionamiento de este algoritmo es necesario modelar su unidad mínima, la neurona. Existen diversos modelos, pero el más famoso y utilizado es el del perceptrón [26].

Un perceptrón es una neurona artificial, y, como se mencionó anteriormente, representa la unidad mínima de una red neuronal. La finalidad de este es de efectuar cálculos para detectar características o tendencias en los datos de entrada.

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

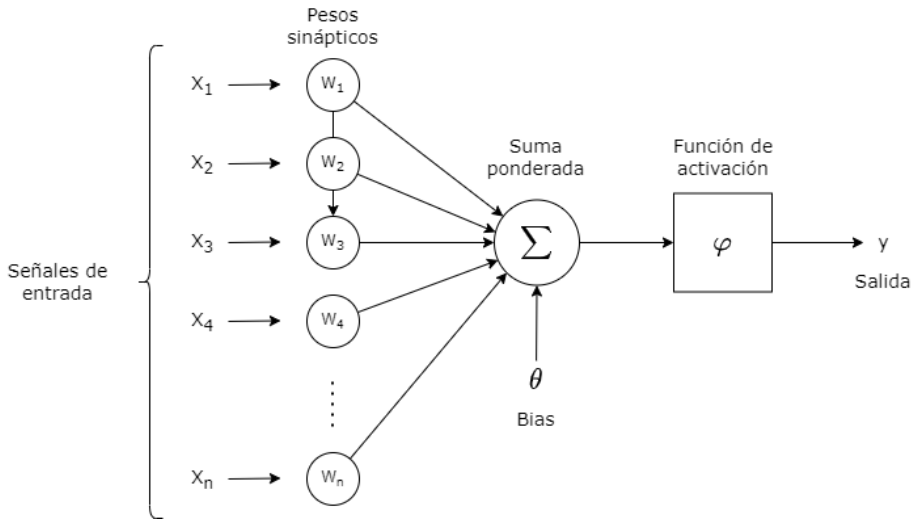

Figura 4. Esquema de una neurona artificial.

De la figura anterior se observa que una neurona artificial cuenta con los siguientes componentes:

- **Entradas:** los datos que van a ser procesados por esta neurona.
- **Pesos sinápticos:** refiere a los parámetros asociados a cada entrada que se ajustan durante cada iteración del entrenamiento. Los mismos buscan ponderar cuál de los parámetros ingresados es el más significativo, es decir, les da un peso.

- **Umbral o Bias:** esta es una entrada externa a la red neuronal y, al igual que los pesos sinápticos, su valor se ajusta durante el entrenamiento.

- **Suma ponderada:** en este bloque se realiza el cálculo de todas las entradas ponderadas. Para ello se calcula el producto interno entre el vector de entradas y el de pesos sinápticos y luego se agrega el valor de bias. Esto cálculo se puede observar en la ecuación 12

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

- **Función de activación:** estas consisten en funciones que se aplican luego de la suma ponderada, y generan la salida de la neurona. Esto cálculo se puede observar en la ecuación 13 y algunos ejemplos de estas se pueden observar en la Figura 5.

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

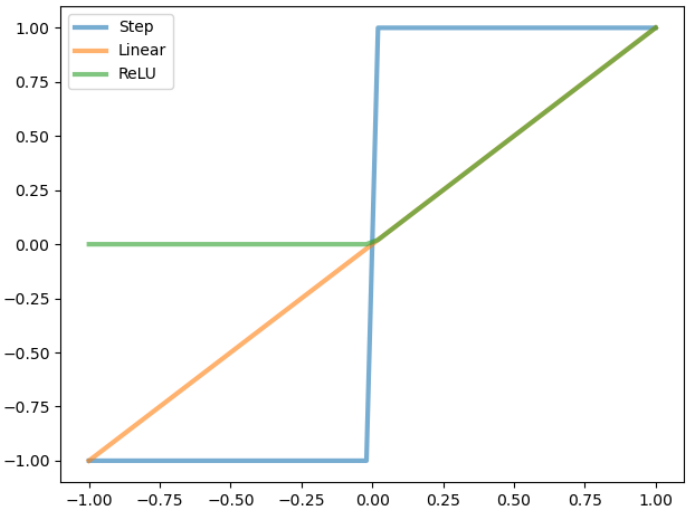

Figura 5. Funciones de activación.

3.6.2. PERCEPTRÓN MULTICAPA

Una vez entendido el funcionamiento de una neurona artificial, el siguiente paso es juntar varias de ellas para generar lo que se conoce como redes neuronales.

Una red neuronal es un conjunto de neuronas interconectadas que se organizan en capas, las cuales poseen una cierta cantidad de estas. Las salidas de las neuronas de una capa suelen constituir las entradas de las neuronas de la capa siguiente. Estos modelos se arman en base a una arquitectura específica, es decir, requieren la definición de la cantidad de capas a utilizar, la cantidad de neuronas, el tipo de estas capas, entre otros parámetros.

Una de las redes neuronales más estudiadas es el modelo del perceptrón multicapa [27]. Este tipo de red cuenta con una capa de entrada que se conecta a una o varias capas ocultas, y una capa de salida. Cada capa puede contener un número distinto de neuronas y se encuentra completamente conectada a la capa adyacente. Cada neurona tiene un peso y un umbral asociados. Si la salida de cualquier nodo (o neurona) individual está por encima del valor de umbral especificado, ese nodo se activa y envía datos a la siguiente capa de la red. De lo contrario, no se pasan datos a la siguiente capa de la red.

Un ejemplo de este tipo de redes se puede observar en la Figura 6.

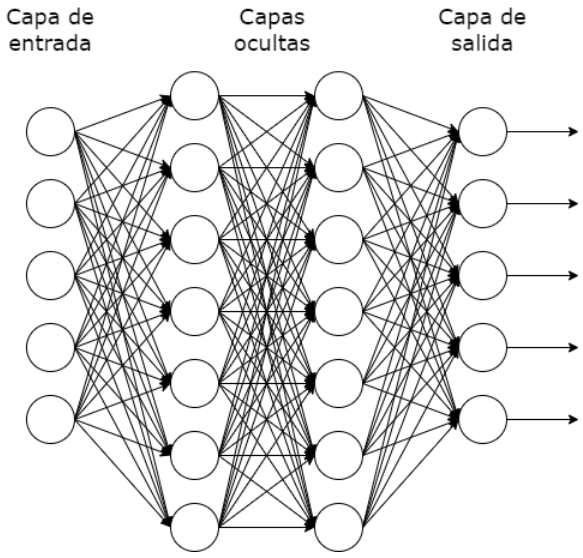

Figura 6. Perceptrón multicapa.

Una de las ventajas más importantes de este algoritmo es que, si la red tiene un número suficiente de neuronas, es capaz de representar cualquier función matemática. Este enunciado se expresa en el teorema de aproximación universal de las redes neuronales [28].

En particular, cuando estas redes neuronales cuentan con dos o más capas ocultas suelen denominarse como redes de aprendizaje profundo [29].

3.6.3. ENTRENAMIENTO DE UNA RED NEURONAL

Los algoritmos de aprendizaje automático se entrenan a partir del procesamiento de datos. En el caso de los modelos de aprendizaje supervisado, los datos de entrenamiento se presentan de a pares (x, y), donde y es el valor objetivo que se espera obtener para el valor de entrada x.

El objetivo de una red neuronal es buscar una configuración de los parámetros entrenables de la red que produzcan que la entrada x genere la salida y. Este proceso se denomina aprendizaje.

Durante el proceso de aprendizaje, se van a producir valores de salida aleatorios que inicialmente van a diferir del valor objetivo. Es por esto que surge la necesidad de tomar un descriptor que sirva para medir esta diferencia. La encargada de medirla es la función de costo. Esta recibe las salidas de la red y las salidas esperadas, y luego calcula una medida de error a partir de una función matemática.

El resultado de esta función se utiliza como una señal de realimentación, para poder ajustar los parámetros entrenables de la red neuronal de manera tal de minimizar el error. Esta tarea es realizada por otra función denominada función de optimización. Esta aplica el algoritmo de propagación del error hacia atrás para calcular el gradiente de la función de costo respecto a los parámetros entrenables de la red. Teniendo en cuenta este gradiente y el valor de tasa de aprendizaje definido en la función de optimización, se puede determinar cómo modificar los parámetros entrenables para lograr disminuir el error de salida.

A las herramientas con las que se cuenta para controlar el proceso de aprendizaje de una red se las diferencian entre dos grupos: parámetros e hiperparámetros. Los parámetros son aquellos valores que se obtienen a través del entrenamiento, mientras que los hiperparámetros se utilizan para parametrizar el propio proceso de instanciación del modelo y sirven como herramientas para describir la configuración del mismo.

Algunos de los hiperparámetros necesarios a definir para el entrenamiento de una red son los tamaños de lotes y las épocas. Los lotes refieren a la segmentación del grupo de datos de entrenamiento. El entrenamiento se realiza en iteraciones, en las cuales en cada una recibe un lote. Una vez procesados todos los lotes, se cumple una época.

Además de los hiperparámetros y los parámetros entrenables, la forma y el tipo de datos que le presentan a la red influyen drásticamente en el desempeño de la misma. Lo

que se busca con este sistema es lograr un grado de generalización que le permita estimar de forma correcta datos que no se le hayan presentado anteriormente. Para conseguir esto, la base de datos se divide en los siguientes conjuntos:

- **Entrenamiento:** este es el conjunto más extenso y se utiliza para la etapa de entrenamiento de la red neuronal.
- **Validación:** este conjunto sirve para medir el desempeño del sistema durante su etapa de entrenamiento.
- **Prueba:** conjunto utilizado para medir el rendimiento a la hora de realizar estimaciones. Consta de datos que nunca fueron vistos por la red neuronal.

Por último, es importante hablar de dos fenómenos indeseados que presentan las redes neuronales cuando no está correctamente implementada, estos son: el sobreajuste y el subajuste. El sobreajuste ocurre cuando el modelo arroja muy buenos resultados con los datos de entrenamiento, pero malos con los de validación y evaluación. Esto implica que el modelo no fue capaz de generalizar el problema, simplemente fue capaz de memorizar los datos de entrenamiento. Por otra parte, el subajuste consta de un error muy grande en la etapa de entrenamiento.

El sobreajuste se da cuando la complejidad del modelo es muy alta en comparación a la tarea que busca resolver o cuando el tamaño del grupo de datos de entrenamiento no fue lo suficientemente grande. Por el contrario, el subajuste se da cuando el modelo es muy simple en comparación a la complejidad de la tarea que intenta resolver.

Para lidiar con un grupo de datos de entrenamiento pobre, una buena técnica a utilizar es la de aumentación de datos. Esta consiste en generar nuevos datos a partir de los ya existentes. Al agregar más casos distintos de prueba en el entrenamiento, la capacidad de generalización del modelo aumenta. Por ende, si nuestra base de datos no es capaz de representar la totalidad de casos del problema a resolver, es posible manipular esta información con el fin de extenderla y generar datos nuevos a partir de los originales.

3.6.4. REDES NEURONALES CONVOLUCIONALES (CNN)

Las redes neuronales convolucionales son un tipo de red en el que las neuronas corresponden a campos receptivos, similar al funcionamiento de la corteza visual. Este algoritmo,

al igual que el de las redes neuronales convencionales, también consta de varias capas de perceptrones, pero a este se le agregan unos filtros que son capaces de detectar patrones en las imágenes [30]. Razón por la cual ganaron gran popularidad en el mundo de la visión artificial.

El bloque básico de las redes neuronales convolucionales es la capa convolucional, la cual tiene las siguientes características:

- **Campo receptivo limitado:** esto refiere a que las neuronas que forman parte de la capa convolucional no están conectadas a todas las entradas como en el caso del perceptrón multicapa. Por el contrario, solo se conectan a una porción de la entrada a la que se domina como campo receptivo. Con esto, es posible modelar estructuras locales.

- **Parámetros compartidos:** en este tipo de redes, los pesos sinápticos de las neuronas se comparten. Esto genera una de sus ventajas más importantes, los patrones aprendidos por sus neuronas son invariantes a la traslación. Por ende, si se aprende un patrón que está ubicado en un lugar específico de la entrada, es posible detectarlo luego en otra ubicación.

- **Convolución:** esta es la característica más importante, la cual le da el nombre a este tipo de red. Consiste en aplicar un filtro de convolución sobre las entradas, la cual, en cada paso, aplica un filtro que realiza la operación de producto interno a una sección de la entrada (la cual suele ser una imagen), cuyo resultado corresponde a un único valor de salida. Las salidas de las capas convolucionales se denominan mapas de características y estos representan la presencia del patrón del filtro al pasar por la señal. En la Figura 7 se puede observar un diagrama de este proceso.

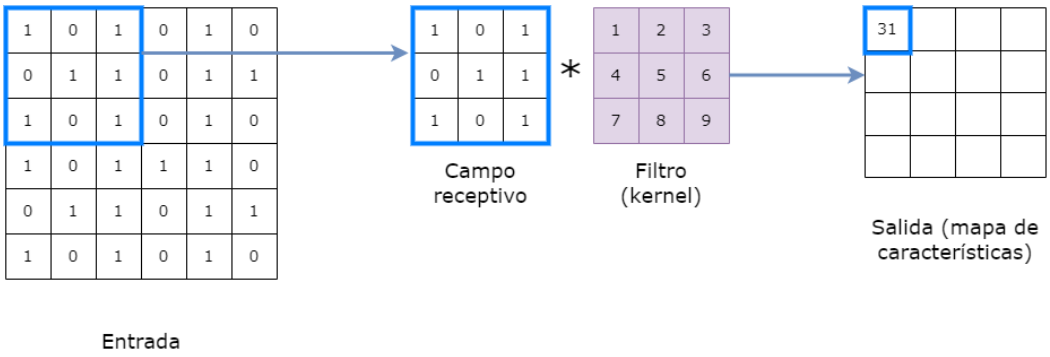

Figura 7. Esquema de un filtro de convolución.

En este tipo de redes se definen otros hiperparámetros extras que los que llevan las de perceptrón multicapa, los cuales son:

- **Tamaño del filtro:** este refiere al tamaño del campo receptivo, el cual suele ser una matriz cuadrada de 3x3 o 5x5.
- **Tamaño del salto:** representa la distancia horizontal y vertical entre campos receptivos de dos unidades contiguas.
- **Relleno de ceros:** cantidad de ceros a agregar en los contornos de la imagen con la finalidad de mantener las dimensiones de entrada y salidas del mismo tamaño.
- **Cantidad de filtros aplicados:** cantidad de filtros convolucionales que se aplican sobre la entrada, la cual es equivalente al número de mapas de características.

En síntesis, este tipo de redes logran generalizar conceptos visuales complejos a partir de una menor cantidad de parámetros que los que usaría una red completamente conectada, los cuales están distribuidos estratégicamente a lo largo de la arquitectura.

4. METODOLOGÍA

4.1. GENERACIÓN DE BASE DE DATOS

Para que una red neuronal sea capaz de aprender una tarea compleja, es necesario contar con una base de datos extensa que sea capaz de representar todas las características del fenómeno que se quiere modelar. A su vez, se debe corroborar que todos los elementos de esta compartan los mismos tipos de codificación, es decir: misma profundidad de bits, frecuencia de muestreo y formato de audio.

Para esta investigación, la base de datos se compone de envolventes temporales de amplitud (TAE) de audios de voz reverberados y los valores de los descriptores acústicos T_{30}, C_{50}, C_{80} y D_{50} de la sala en la que se reverberó el audio en cuestión. Se tomó la decisión de guardar las TAE en lugar de los audios reverberados para disminuir drásticamente el peso de la base, ya que cada TAE se almacena como una lista de 200 muestras mientras que los audios ocupan 80000.

Para generarla, se utilizaron audios de voz anecoicos grabados en la cámara anecoica de la Universidad Tecnológica de Delft [31], los cuales se reverberaron mediante su convolución con una base de respuestas al impulso de salas. Con lo cual, teniendo los audios reverberados y la respuesta al impulso, se calcularon los valores de TAE con el audio en cuestión y los descriptores acústicos con la respuesta al impulso que lo reverberó.

Una vez generada la base de datos, la misma se divide en tres grupos: conjunto de entrenamiento, validación y prueba.

4.2. AUMENTACIÓN DE RESPUESTAS AL IMPULSO

Para que una red neuronal sea capaz de aprender a resolver una tarea compleja, es necesario contar con una base de datos que pueda representar la totalidad de casos del fenómeno a estimar. En el caso del conjunto de datos en esta investigación en particular, se necesitan de muchas respuesta al impulso cuyo tiempo de reverberación y su pendiente de decaimiento representen la mayor cantidad de salas posibles, lo cual resulta muy complejo y costoso obtener mediante mediciones concretas.

Una forma de resolver esta problemática es mediante técnicas de aumentación de datos. Esto consiste en realizar un tratamiento en la señal para alterar los parámetros acústicos

de cada respuesta al impulso. En concreto, se analizarán técnicas para modificar los descriptores acústicos T_{60} y DRR, de forma tal de podes ampliar considerablemente la cantidad de salas modeladas en el conjunto de datos.

Es importante aclarar que el conjunto de respuestas aumentadas, en esta investigación, se consigue a partir del tratamiento sobre señales de impulsos reales. Para esto, se utilizan dos procesos: una modificación de amplitud en la parte temprana de la respuesta al impulso para controlar la relación directo-reverberado y una alteración de envolvente de caída para controlar el tiempo de reverberación [17].

4.2.1. AUMENTACIÓN DE LA RELACIÓN DIRECTO-REVERBERADO

Según lo estudiado en la sección 3.4, se sabe que el descriptor DRR divida la respuesta al impulso en dos partes: la que proviene del sonido directo y la del campo reverberado. A la primera se la denotará como $h_e(t)$ mientras que a la segunda como $h_l(t)$.

Para aumentar este descriptor, a la parte correspondiente al sonido directo (la cual, por convención, suele tomarse como los primeros 2.5 ms de la respuesta) se le aplica una ganancia determinada por un factor α, la cual se ajusta para obtener el valor de DRR deseado. Esto genera una nueva señal denominada $\tilde{h}_e(t)$. Este proceso por sí solo genera discontinuidades en la señal, para evitarlo es necesario aplicar ventanas complementarias. Al hacerlo, se obtiene una señal directa ventaneada y un residuo ventaneado. Este cálculo se realiza a partir de lo expresado en la ecuación 4.2.1.

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

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

Ahora, combinando lo expresado en la ecuación y 11 se plantea un sistema de ecuaciones en donde se define un valor deseado de DRR y se despeja el valor de α para obtenerlo.

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

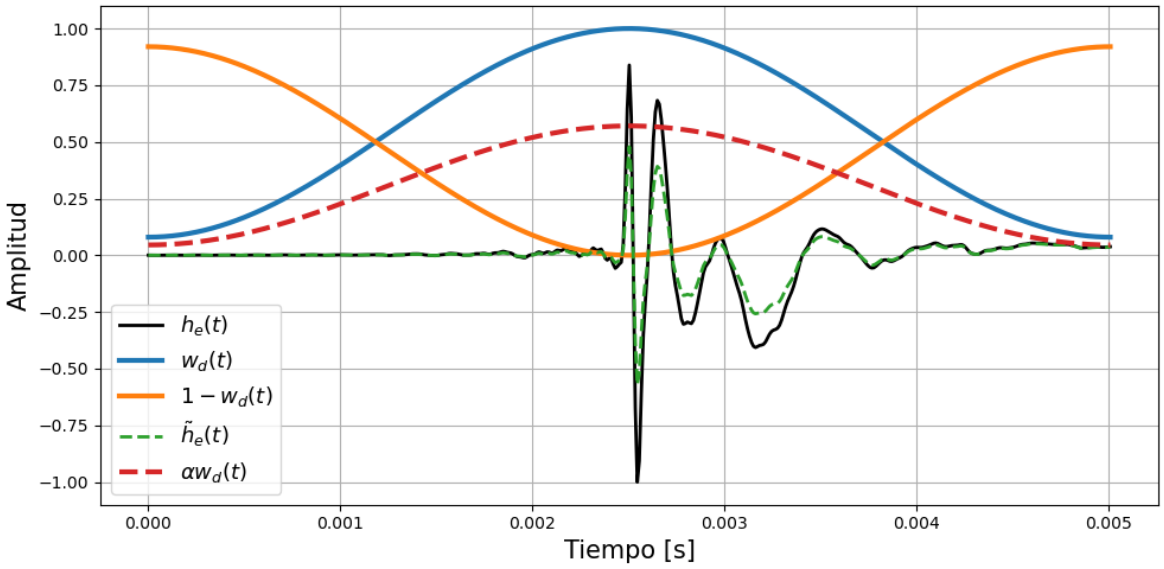

Figura 8. Proceso de aumentación del DRR.

Por último, la parte directa generada se concatena con el resto de la respuesta al impulso, generando la nueva respuesta aumentada.

Es importante aclarar que, para tiempos de reverberación muy cortos, no siempre es posible generar relaciones directo-reverberado muy bajas, esto debido a que la energía tardía ya es demasiado baja desde un principio. Para solucionar esto, se definen valores límites para el valor de α aplicado.

4.2.2. AUMENTACIÓN DEL TIEMPO DE REVERBERACIÓN

A diferencia del caso del DRR, para la aumentación del descriptor T_{60} se trabaja con la parte tardía de la respuesta al impulso.

Como se estudió en la sección 3.3, la RIR se puede modelar como un ruido blanco Gaussiano con una caída de nivel dominada por una exponencial decreciente. A su vez, es importante mencionar que la caída de la exponencial es dependiente de la frecuencia y, también, se asocia a esta un determinado piso de ruido.

Teniendo en cuenta esto, se considera el modelo de síntesis de respuestas al impulso de la ecuación 15 para el caso de estudio de aumentación.

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

Donde A_m es la amplitud inicial, τ_m la tasa de caída, σ_m el nivel de piso de ruido, $n(t)$

corresponde al ruido Gaussiano de media cero y desvío estándar uno, t_0 el tiempo donde comienza la parte tardía de la RIR, m el índice que indica la sub-banda con la que se está trabajando y $u(t)$ un escalón unitario.

Para este modelo, el tiempo de reverberación se relaciona con el parámetro τ según lo expresado en la ecuación 16.

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

Donde T_s representa el período de muestreo.

Al modelo de RIR la ecuación 15 se le aplican métodos de optimización no lineales para estimar los parámetros $\hat{A}_m$, $\hat{\tau}_m$ y $\hat{\sigma}_m$ que mejor aproximen la envolvente de caída de la respuesta al impulso. Con estos y con el cálculo de la tasa de caída deseada $\tau_{m,d}$ determinada a partir del tiempo de reverberación buscado, se modifica la parte tardía de la RIR inicial multiplicándola con una envolvente exponencial creciente o decreciente según corresponda. Esta modificación se puede observar en la ecuación 17.

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

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

El proceso de aumentación de tiempo de reverberación consiste, básicamente, en modificar la pendiente de caída de la RIR hasta obtener una deseada para cada banda frecuencial. Una vez obtenidas, las sub-bandas generadas se suman para obtener la respuesta final.

El proceso anteriormente descripto funciona solamente si se busca sintetizar tiempos de reverberación menores al de la señal original, es decir, solo en los casos en los que se multiplica la respuesta al impulso por una exponencial decreciente. Cuando se quieren generar tiempos de reverberación mayores, es necesario multiplicar por una exponencial creciente, lo que produce una amplificación de la parte tardía de la RIR. Al hacer esto, se amplifica también el piso de ruido presente en la señal, lo que produce pendientes de caída inestables que no se corresponden con la respuesta original.

Para evitar este problema, es necesario estimar el piso de ruido de la respuesta al impulso. En concreto, se utiliza el método de Lundeby [22] para esta tarea. Una vez estimado, la respuesta final se obtiene haciendo un cross-fade en el inicio del piso de ruido

entre la parte tardía generada y una cola reverberante sintética, la cual se obtiene a partir de multiplicar ruido Gaussiano con una envolvente exponencial decreciente, utilizando los parámetros previamente calculados.

A modo de resumen, los pasos a seguir para realizar la aumentación del tiempo de reverberación son:

- Normalización de la respuesta al impulso.
- Filtrado por bandas de octava.
- Estimación del piso de ruido.
- Estimación de la envolvente de caída.
- Sintezis de una señal utilizando la envolvente estimada con piso de ruido nulo a una señal de ruido Gaussiano.
- Cross-fade entre la señal sintetizada y la original en el punto donde inicia el piso de ruido.
- Multiplicar la señal por una exponencial creciente/decreciente.
- Suma de las sub-bandas para obtener la RIR sintetizada en su espectro completo.
- Combinar la parte tardía aumentada con la directa de la respuesta al impulso original.

4.3. CURVA DE DECAIMIENTO DE UNA SEÑAL DE VOZ

Retomando la idea de pensar a una sala con un micrófono y una fuente dentro como un sistema, es posible entender que toda señal emitida en su interior se va a ver afectada por el mismo tal como se observa en la ecuación 1. Visto de otra manera, la señal captada en el interior del recinto va a contener información del mismo.

Sabiendo esto, en la investigación de Kendrick et al. [10] observaron que la curva de decaimiento del final de una palabra tiene cierta similitud con la de una RIR. A su vez, esta curva se ve afectada por la reverberación de la propia sala, haciendo que el tiempo de decaimiento sea mayor debido al tiempo de reverberación del recinto.

Este fenómeno es muy importante, ya que al grabar una oración dentro del recinto, los investigadores fueron capaces de determinar el tiempo de reverberación de la sala a

partir de calcular la media de la distribución de los tiempos de reverberación obtenidos de las diferentes curvas de decaimiento que se pueden extraer del audio captado. Esta fue la primera técnica utilizada para estimar el tiempo de reverberación de un recinto de forma ciega.

A modo ilustrativo, en la Figura 9 se compara la curva de decaimiento de una señal contra sí misma pero afectada por la reverberación de un recinto.

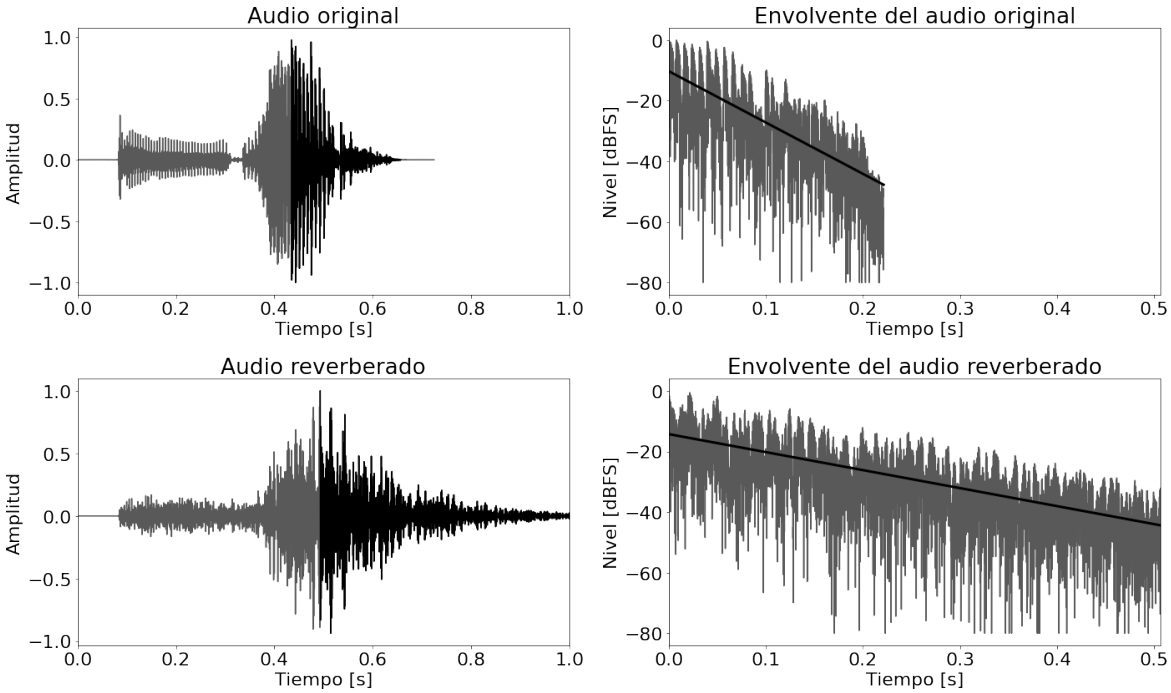

Figura 9. Diferencias en la curva de decaimiento para un audio limpio contra uno reverberado.

4.4. OBTENCIÓN DE RESPUESTAS AL IMPULSO

En esta investigación se utilizaron respuestas al impulso reales y sintéticas. Además, se genera un tercer grupo conformado a partir de la aumentación de las RIR reales.

En todos los casos, las señales cuentan con una frecuencia de muestreo de 16 kHz. Esto es importante porque se busca homogeneizar la codificación de la base de datos para evitar problemas a la hora de entrenar la red neuronal.

4.4.1. RESPUESTAS AL IMPULSO SINTÉTICAS

Para generar los impulsos sintéticos, se utilizó el modelo de Schroeder expresado en la ecuación 7. Este permite sintetizar RIRs de salas paralelepipédicas dado un determinado tiempo de reverberación. Para ello, la señal se genera a través de la multiplicación de

un ruido blanco gaussiano de media cero y desvío estándar unitario con una exponencial decreciente.

En concreto, se generaron de forma equitativa un total de 2900 respuestas al impulso, cuyos valores de tiempo de reverberación van desde $0{,}2s$ a $3s$, con pasos de $0{,}1s$. Se tomó este rango para respetar el mismo que se utilizó en el trabajo [9]. En cuanto a la duración de las señales, se decidió que el largo de las mismas sea de $0{,}5s$ más que el tiempo de reverberación con el que se sintetizaron.

A modo ilustrativo, en la Figura 10 se pueden observar 3 respuestas al impulso generadas utilizando el método de Schroeder con distinto tiempo de reverberación.

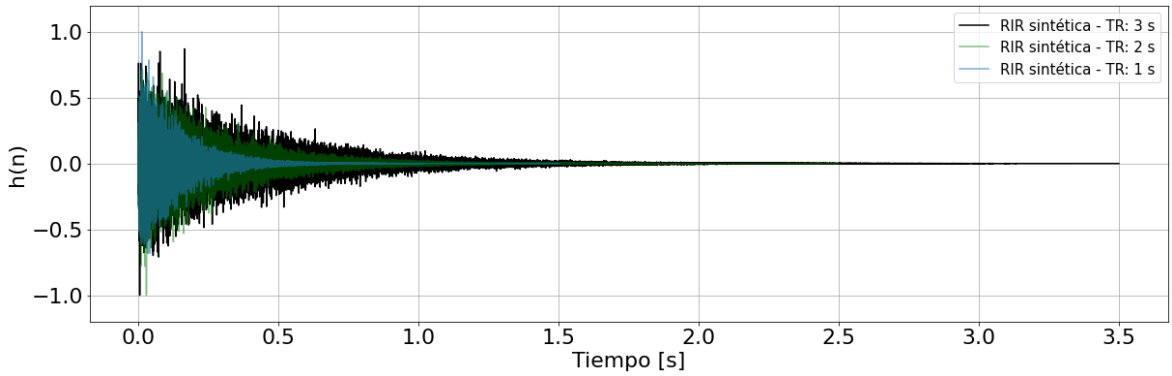

Figura 10. Respuestas al impulso con distintos tiempos de reverberación generadas de forma sintética.

4.4.2. RESPUESTAS AL IMPULSO REALES

En cuanto a las respuestas al impulso reales, estas se obtuvieron de la base de datos C4DM [32].

Estas RIRs se midieron utilizando la téncina del barrido frecuencial (LSS) [18]. Para generar la señal, se usó como fuente un altoparlante Genelec 8250A y como micrófono de medición uno omnidireccional modelo DPA 4006. El altoparlante mencionado consiste en un transductor de dos vías, siendo la de frecuencias bajas y medias un driver de 8’ y uno de 1’ para las frecuencias altas.

En concreto, se midieron tres recintos para generar este conjunto de datos. El primero consiste en una sala multipropósito con, aproximadamente, 800 asientos (*Great Hall*). El segundo se trata de una biblioteca con un estilo victoriano (*Octagon*). Por último, la tercera sala es 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 fuente y diferentes posiciones de micrófono. Estos se ordenaron de forma tal de formar una grilla equiespaciada dentro de las salas, generando así un mapeo uniforme del recinto.

En la Tabla 1 se pueden observar los tiempos de reverberación para 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 pueden observar los esquemas de medición de las salas medidas para generar la base de datos.

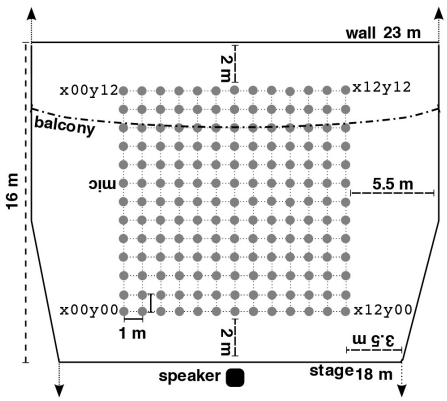

Figura 11. Esquema de medición de respuestas al impulso del recinto Great Hall.

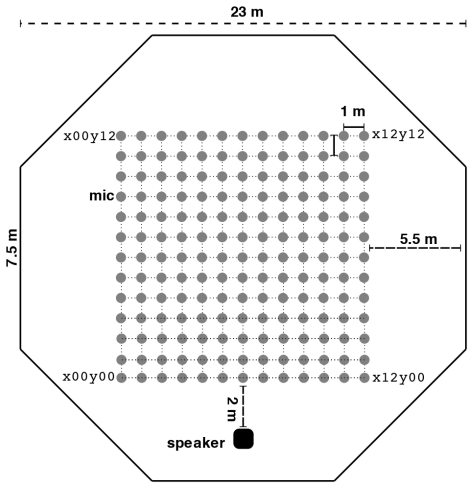

Figura 12. Esquema de medición de respuestas al impulso del recinto Octagon.

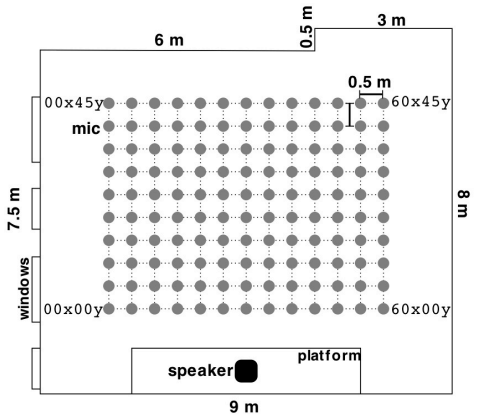

Figura 13. Esquema de medición de respuestas al impulso del recinto Classroom.

4.4.3. RESPUESTAS AL IMPULSO AUMENTADAS

Por último, el conjunto de RIRs aumentadas se obtuvo a partir de procesar los impulsos reales de la base de datos anteriormente descripta.

Para esta tarea, se eligieron, de manera aleatoria, 15 respuestas al impulso por cada sala, dando un total de 45 a procesar. A estas se les aplicaron las técnicas de aumentación descriptas en el capítulo 4.2 para variar sus parámetros T_{60} y DRR.

En concreto, para el caso del descriptor T_{60}, se procesó cada RIR de forma tal de generar respuestas nuevas de cada una, cuyos valores van desde 0.2 a 3 s en pasos de 0.1 s. En paralelo, se realiza otro proceso de aumentación para el DRR, el cual consiste en seleccionar aleatoriamente 5 de las señales aumentadas en el paso anterior y obtener de estas nuevas señales cuyas relaciones directo-reverberado van desde los -6 a 18 dB en pasos de 1 dB.

Por lo tanto, por cada RIR seleccionada, se realiza un total de 29 aumentaciones para variar su tiempo de reverberación y 125 para modificar su relación directo-reverberado. Con esto, se procesan un total de 6930 respuestas al impulso.

Es importante aclarar que algunas de las señales a las que se les quiere aumentar el tiempo de reverberación no pueden ser procesadas debido al nivel de ruido intrínseco que presentan las mismas en ciertas bandas de frecuencia. En estos casos, la respuesta en esa banda se descarta de la base de datos.

Para realizar el análisis por banda para aumentar el tiempo de reverberación, se utilizó un banco de filtros de bandas de octava como el que se observa en la Figura 14. Para las bandas entre 125 y 4000 Hz se utilizaron filtros pasa banda de tipo butterworth de orden

4 con sus frecuencias de corte normalizadas y para la banda de 8000 Hz se utilizó un filtro pasa altos con las mismas características.

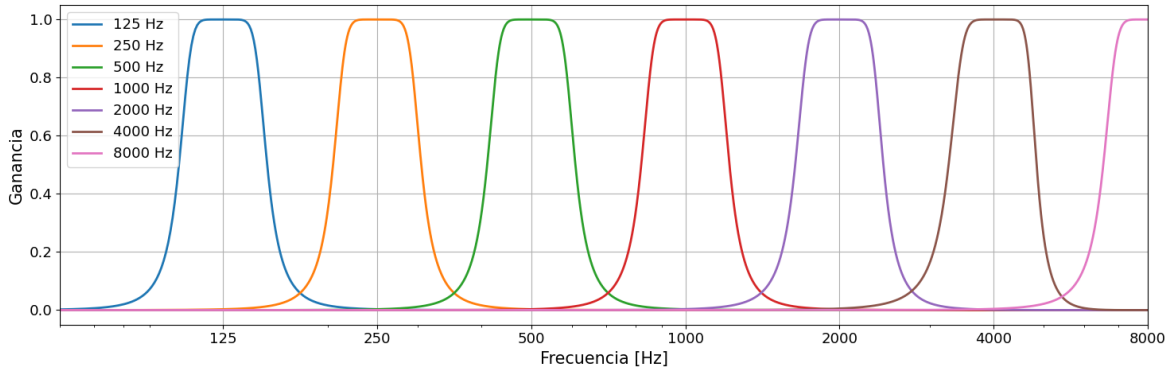

Figura 14. Banco de filtros para el análisis de aumentación de tiempo de reverberación de las respuestas al impulso.

4.5. BASE DE DATOS DE SEÑALES DE VOZ

Como se describió anteriormente, las señales reverberadas se obtienen a partir de la convolución entre una señal de habla anecoica y la respuesta al impulso de una sala. Por tanto, una vez generadas las RIRs, es necesario conseguir una base de datos de señales de voz.

En concreto, para esta investigación se utilizó la base de datos generada para el desafío *Acoustic Characterisation of Environments Challenge* (ACE Challenge) [13]. La misma cuenta con un conjunto de respuestas al impulso grabadas en 5 salas y audios de voz anecoicos. En cuanto a las señales de habla, se cuenta con un total de 50 audios de grabaciones de personas respondiendo una pregunta con una oración extensa, de forma tal de generar dinámica en la señal mediante los silencios entre cada palabra. Para esto, se entrevistaron a 5 hombres y 5 mujeres que responden a 5 preguntas, las cuales se separan en los distintos archivos.

De todo este conjunto, se descartaron 12 audios ya que su duración era inferior a 5 s y se seleccionaron los 38 restantes. Luego, se tomaron de forma aleatoria 30 de estos archivos para conformar la base de datos para el entrenamiento y los 8 restantes se separaron para el conjunto de pruebas.

Una vez seleccionados, se recortaron los archivos de audio para que tengan una duración total de 5 s y se remuestrearon para que se frecuencia de muestreo sea de 16 kHZ.

4.6. CÁLCULO DE LOS DESCRIPTORES DE LAS RESPUESTAS AL IMPULSO

Después de obtener la base de datos de respuestas al impulso, el siguiente paso es calcular los descriptores acústicos T_{30}, C_{50}, C_{80} y D_{50} para cada señal. Para esto, se utilizan los lineamientos de la norma ISO 3382 descriptos en la sección 3.2.

En todos los casos, las RIRs se filtraron usando el banco de filtros de la Figura 14 y posteriormente se calcularon los parámetros por bandas de frecuencia.

Debido a que algunas bandas pueden tener mucho ruido intrínseco en el audio, se decidió estimar el piso de ruido de las señales utilizando el método de Lundeby a la hora de calcular el tiempo de reverberación. Además, se tomó como criterio de aceptación que las bandas tengan un rango dinámico de, por lo menos, 45 dB desde el punto máximo de la respuesta hasta su piso de ruido estimado. En todos los casos que este criterio no se cumple, se rechaza esta banda y se continúa con las demás.

Por tanto, la cantidad de descriptores con los que se cuenta por cada banda de frecuencia queda delimitado exclusivamente por el nivel de piso de ruido de las señales en esa banda en cuestión.

4.7. OBTENCIÓN DE LAS ENVOLVENTES TEMPORALES DE AMPLITUD: TAE

4.7.1. GENERACIÓN DE TAE SIN RUIDO

Una vez generadas las bases de datos de señales de habla y de respuestas al impulso, se procede a convolucionar uno a uno los datos y para obtener los audios reveberados. Luego, de estos, se calculan las envolventes temporales de amplitud.

Para obtener las TAE se utiliza el proceso detallado en la sección 3.5. Como primer paso, se convoluciona la señal de habla con una RIR, pero quedándose solo con 5 s de este audio generado y no el total de la convolución. Luego, se filtra la señal con el mismo banco de filtros que se utilizó para la aumentación del tiempo de reverberación, el cual se puede ver en la Figura 14. Por lo tanto, se obtienen 7 versiones de la señal, uno por cada banda en cuestión. A cada una de estas, se le aplica la transformada de Hilbert para obtener la envolvente de las señales y luego se les aplica un filtro pasa bajos con frecuencia de corte en 20 Hz para suavizar aún más la curva. Luego, se remuestrea cada señal a una frecuencia de muestreo de 40 Hz para reducir la tasa de información y aliviar el proceso de entrenamiento para la red neuronal. Finalmente, se normaliza cada señal para que sus valores de amplitud

estén entre 0 y 1.

Con todo este proceso, se logra comprimir considerablemente la información, pasando de señales de 80000 muestras a, solamente, 200. Además, otra cosa importante a destacar es que todo el contenido de las palabras se perdió por completo, quedando solamente la información de su dinámica como se observa en la Figura 3.

Por consecuencia, para que la red neuronal pueda extraer los parámetros acústicos de estas señales, las mismas deben contar con mucha dinámica ya que va a extraer los mismos a partir de las curvas de decaimiento presentes en el audio. En los casos de las grabaciones anecoicas esto es sencillo ya que simplemente hay que cuidar que el piso de ruido de la sala de grabación sea bajo, pero esto no es representativo de un caso real para el que se busca entrenar este modelo. Esto se debe a que cuando se graba un audio reverberado en una sala real, el mismo va a contar con cierto nivel de ruido propio del lugar el cual no siempre es posible de minimizar y puede ser muy elevado, con la posibilidad de que estén por sobre los valores propuestos por la norma para la medición de respuestas al impulso.

A saber, los ruidos que pueden perjudicar la dinámica de la envolvente temporal de amplitud son los monótonos y constantes. Por ejemplo, el ruido de las aspas de un ventilador encendido, el murmullo de la gente, un flujo vehicular constante, entre otros. Todos los demás tipos de ruido, como los impulsivos, no afectan al análisis de la TAE ya que al pasar por todo el proceso de su obtención, los mismos se reflejan como subidas y bajadas de nivel, lo cual ayuda con la dinámica del audio. Algunos ejemplos de estos podrían ser golpes de puerta, caída de un objeto, entre otros.

Debido a esto, se decidió generar una segunda base de datos de envolventes temporales de amplitud agregándole ruido constante a las señales con el fin de comparar las predicciones de los modelos entrenados con los distintos conjuntos.

4.7.2. GENERACIÓN DE TAE CON RUIDO ROSA

Para este segundo conjunto de datos, el proceso de obtención de las TAE es prácticamente el mismo. En concreto, el único paso que difiere es que, una vez filtradas las señales reverberadas, se les agrega un ruido controlado de forma tal de conseguir un cierto nivel de relación señal-ruido (SNR).

Como se describió anteriormente, los ruidos que complejizan la tarea de estimación

para la red neuronal son los de tipo monótono y constante. Por lo tanto, se decidió agregarles ruido rosa a las señales ya que son los que presentan mayor dificultad para el algoritmo.

Este ruido fue agregado de forma aleatoria siguiendo una distribución uniforme, con la finalidad de que la señal obtenida tenga un SNR de 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 en cuestión. A saber, el parámetro SNR se define según la ecuación 18.

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

Por lo tanto, si se conoce el valor RMS de la señal y del ruido, es posible multiplicar a este último por un escalar para modificar su amplitud y así obtener el valor de SNR deseado. Este escalar se puede obtener a partir de la ecuación 19.

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

Donde a representa el factor de compensación de la señal de ruido, $SNRrequerido$ es el valor de relación señal-ruido buscado, RMS_{ruido} es el valor de RMS del ruido sin compensar y $RMS_{señal}$ es el valor RMS de la señal de voz.

Por último, la señal reverberada con el nivel de SNR buscado se obtiene según la ecuación 20.

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

4.8. MODELO PROPUESTO

Recordando, el objetivo de esta investigación es entrenar una red neuronal que sea capaz de estimar los parámetros acústicos de una sala a partir de un audio de voz reverberados. Para lograr esto, se propone un método inspirado en la combinación de dos trabajos dentro de este campo. Se utilizan los lineamientos definidos en la investigación de Kendrick et al. [10] sobre la obtención del tiempo de reverberación de una sala a partir de la curva de decaimiento de un audio reverberado y una variación de la arquitectura de red neuronal del modelo propuesto en la investigación de Duangpummet et. al. [9] para la estimación del tiempo de reverberación a partir de una señal de habla reverberada.

En concreto, se plantea un sistema como el que se puede observar en la Figura 15 para realizar esta tarea. Como entrada al mismo se utiliza una señal de habla reverberada de 5 s de duración y 16 kHz de frecuencia de muestreo. A este audio se lo pasa por el banco de filtros de la Figura 14 y se obtienen 7 versiones del mismo (una por cada banda). En este punto la tarea se paraleliza, analizando cada banda por subsistemas idénticos. Como primer paso de este, se encuentran las envolventes temporales de amplitud utilizando el método detallado en la sección 4.7.1. Una vez generada, esta pasa como parámetro de entrada de una red neuronal convolucional (CNN) y a la salida se obtiene la estimación de los descriptores T_{30}, C_{50}, C_{80} y D_{50}. El sistema cuenta con 7 de estas redes con arquitecturas idénticas, variando únicamente en que en su etapa de entrenamiento se usaron valores de TAE filtrados por bandas distintas.

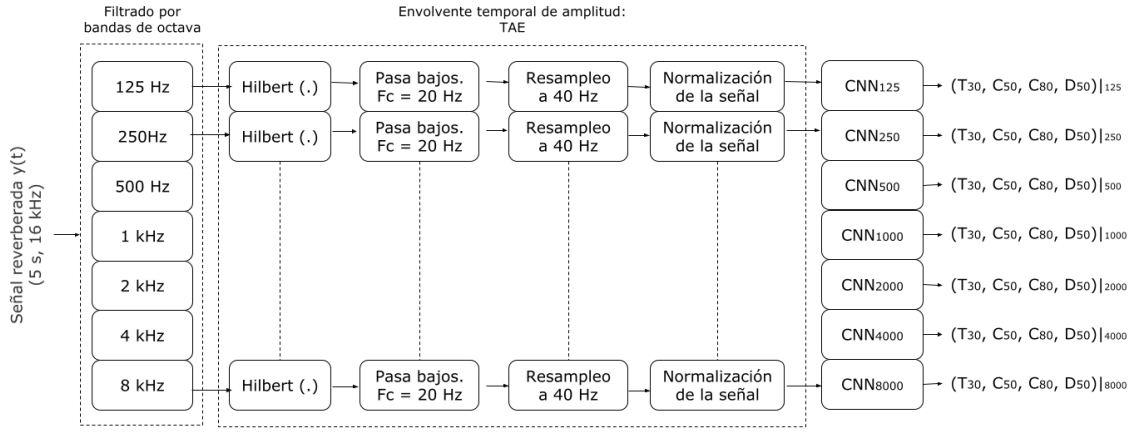

Figura 15. Diagrama en bloques del modelo propuesto.

En cuanto a la red neuronal, como se mencionó en el párrafo anterior, se crearon 7 CNN de una única dimensión, cuya función es estimar los parámetros acústicos a partir de las características de las TAEs de entrada. Estas redes son idénticas y se utiliza una por cada banda.

Cada una de estas cuenta con 4 capas convolucionales. La primera, denominada de entrada, toma los valores de la TAE a ser convolucionada con los filtros. En cada una de estas capas se utiliza una activación no lineal de tipo ReLu. Entre cada capa convolucional se aplica una reducción de dimensiones a través de un filtrado por max pooling y luego realiza una normalización por lotes (batch normalization). La tasa de abandono antes de la última

capa se establece en 40% para evitar que la red memorice los datos. Finalmente, la capa completamente conectada es la de salida, la cual trata de estimar los descriptores acústicos a través de conocer su valor esperado, utilizando como función de costo el error cuadrático medio (MSE) entre el valor obtenido y el esperado. La compilación del modelo se hace a través del optimizador de Adam [33] con un valor de tasa de aprendizaje de 0.001.

En la tabla 2 se puede observar la arquitectura de la red propuesta para esta investigación.

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)

A saber, se entrenaron dos de estos sistemas. La diferencia entre ellos son los pares de datos de entrada/salida que se utilizan en la etapa de entrenamiento. En el primero, los valores de entrada son las envolventes temporales de amplitud sin ruido agregado y en el segundo se usa la base de datos con el ruido rosa añadido. En ambos casos, los datos de salida son los parámetros acústicos que se intentan estimar. La necesidad de generar estas dos versiones se detalla con profundidad en la sección 4.7.

La arquitectura de estas redes neuronales se implementaron utilizando la biblioteca Tensorflow [34] en el lenguaje de programación python.

4.9. EVALUACIÓN DEL MODELO

Para evaluar el desempeño de la redes, una vez entrenadas, se utilizan los pesos sinápticos obtenidos en cada capa de los modelos y se realizan las predicciones de los parámetros

acústicos a partir de una TAE desconocida, comparándolos con los valores reales. Para el primer modelo, la TAE no tendrá ruido añadido mientras que para el segundo sí cuenta con el ruido rosa agregado.

Este proceso se realiza para cada modelo con todo el conjunto de datos de pruebas para cada caso y se calculan los coeficientes de correlación de Spearman [35] para determinar qué tan eficientes son para estimar los valores. Esto se repite para las 7 bandas de estudio.

Finalmente, se comparan los valores de correlación obtenidos en cada modelo para determinar si la presencia de ruido en las señales perjudica en el entrenamiento del sistema y la estimación de los parámetros.

4.10. COMPARACIÓN DEL MODELO CON MEDICIONES DE CAMPO

Finalmente, como último medio de comparación, se realizaron mediciones de campo de respuestas al impulso de tres salas siguiendo los lineamientos de la norma ISO 3382 y el método del barrido frecuencial. En cada una de estas, se tomó una posición de fuente y dos de micrófonos para poder comparar los resultados de parámetros acústicos obtenidos con cada uno.

A su vez, usando el mismo equipo y las mismas posiciones de micrófonos y de fuente, se grabaron audios de habla de 5 segundos de duración reproducidas por el parlante con la finalidad de que las tomas cuenten con la reverberación del recinto en cuestión. La decisión de utilizar audios de voz pregrabados en vez de una persona hablando en vivo se tomó por dos motivos: en principio, esta era la única forma de asegurar que la posición de la fuente era la misma en ambos métodos (esto es importante ya que los parámetros acústicos de la sala varían cuando se cambia de posición la fuente y/o el micrófono) y, por otro lado, reproducirlo en parlantes permite controlar el nivel de la señal.

Los audios de voz capturados se procesaron para obtener sus respectivos TAEs y, posteriormente, se los pasó por la red entrenada sin ruido en la base de datos para poder estimar los parámetros acústicos del recinto y luego compararlos con los obtenidos con el método normado de las respuestas al impulso. Se decidió utilizar solo esta red ya que las condiciones de ruido en el lugar no eran muy altas, por ende el modelo que se entrenó utilizando ruido en la base de datos quedaba en desventaja.

Los instrumentos utilizados para las mediciones fueron dos micrófonos Behringer ECM8000 [36] y un parlante Sony MHC-EC99 [37], el cual emitió un barrido frecuencial de 20 a 13k Hz para encontrar las respuestas al impulso de cada recinto. Este rango de frecuencias fue elegido de forma tal de poder tener información en las frecuencias de octavas que se querían estudiar, las cuales van de 125 a 8k Hz. Por otra parte, en la Figura 16 se puede observar una imagen de una de las salas medidas y, además, los instrumentos utilizados.

Figura 16. Imagen de una de las salas medidas.

Por último, en las Figras 17, 18 y 19 se pueden observar los esquemas de las 3 salas utilizadas durante las mediciones.

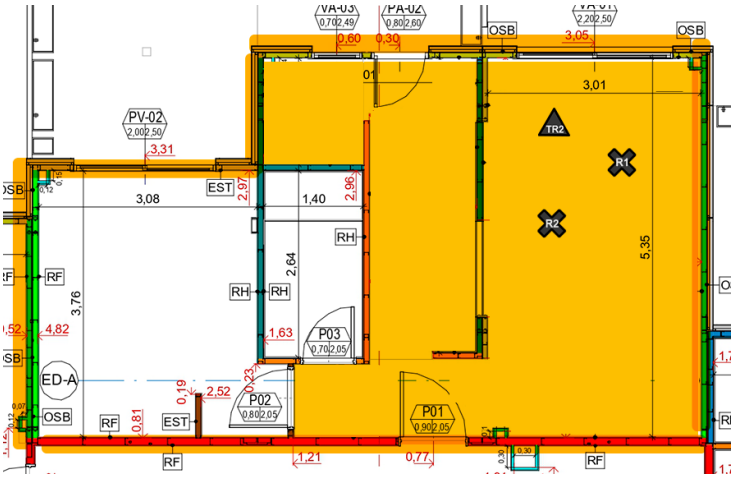

Figura 17. Esquema de medición de la sala 1.

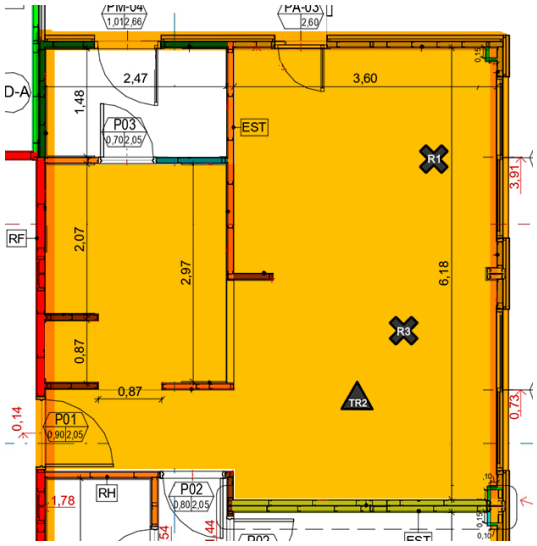

Figura 18. Esquema de medición de la sala 2.

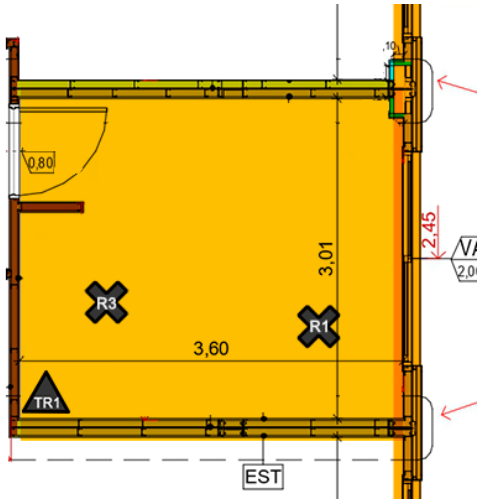

Figura 19. Esquema de medición de la sala 3.

5. RESULTADOS Y ANÁLISIS

El código desarrollado para esta investigación, y su correspondiente documentación, se encuentran disponibles en un repositorio público de Github [38]. En él, se pueden observar las implementaciones de las redes neuronales, los algoritmos para el cálculo de los descriptores acústicos y el análisis de los datos obtenidos.

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

Cómo se mencionó en la sección anterior, para que la red neuronal sea capaz de generalizar el aprendizaje y así poder estimar los parámetros acústicos de salas desconocida, es necesario contar con una base de datos que sea capaz de representar la totalidad de casos en una medición real.

En el contexto de esta investigación, esto corresponde a tener una representación de la mayor cantidad de recintos posibles. Para esto, una forma de estudiar la versatilidad de salas con las que se cuenta en el conjunto de datos es mediante el análisis de los tiempos de reverberación y la relación directo-reverberado de las RIRs.

Se decidió estudiar el $T30_{mid}$ y el DRR_{mid} de las señales. Estos descriptores surgen a partir del promedio sus valores en las bandas de 500 y 1000 Hz.

En la Figura 20 se pueden observar los valores de los descriptores $T30_{mid}$ y DRR_{mid} para las respuestas generadas de forma sintética.

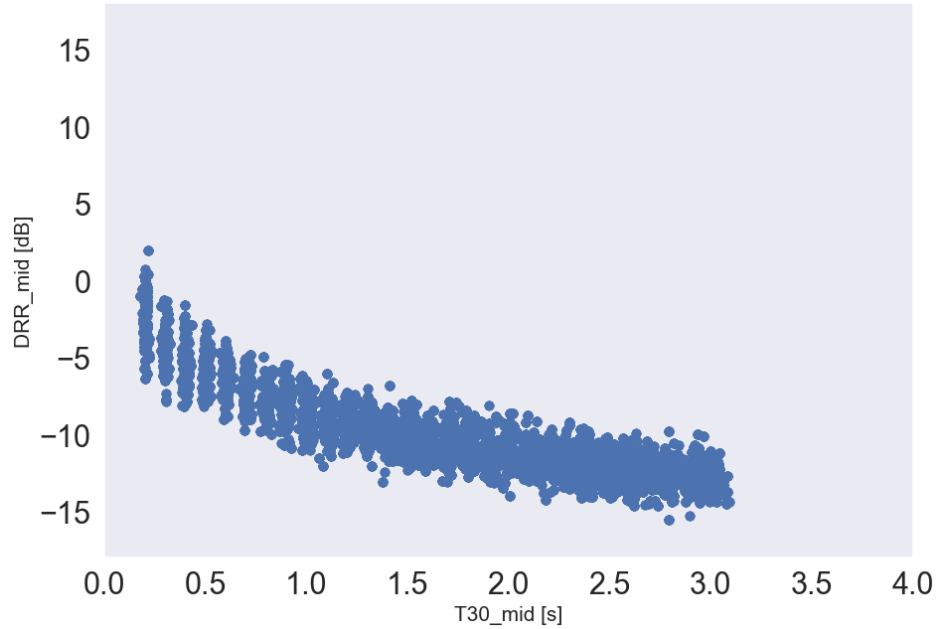

Figura 20. Gráfico de dispersión del $T30_{mid}$ contra el DRR_{mid} para las RIR sintéticas.

En este conjunto de la base de datos se puede observar cómo disminuye el valor del DRR a medida que el tiempo de reverberación aumenta en las señales. Esto permite determinar que, al menos utilizando este método de sintetización de respuestas al impulso, no es posible contemplar todo el conjunto de recintos que se pueden encontrar en situaciones de medición reales. Para cambiar esto, habría que utilizar alguna técnica que permita manipular el valor del DRR dejando el tiempo de reverberación intacto durante la síntesis.

El segundo grupo a analizar es el de las respuestas reales de la base de datos, el cual se puede observar en la Figura 21.

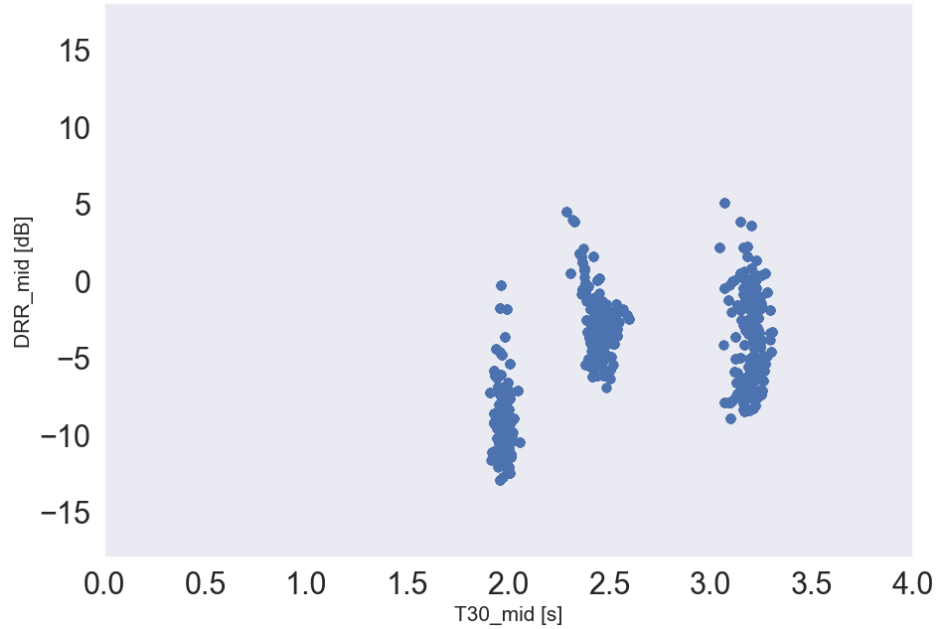

Figura 21. Gráfico de dispersión del $T30mid$ contra el DRR_{mid} para las RIR de salas reales.

En este gráfico se pueden observar claramente los 3 recintos medidos, siendo el primer cúmulo de puntos el correspondiente al *classroom*, seguido por el *great hall* y luego, con el mayor tiempo de reverberación, el *octagon*.

Aún juntando ambos conjuntos, se hace notoria la falta de respuestas que simulen valores altos y bajos de DRR, lo cual realza la necesidad de utilizar técnicas de aumentación para obtenerlas.

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

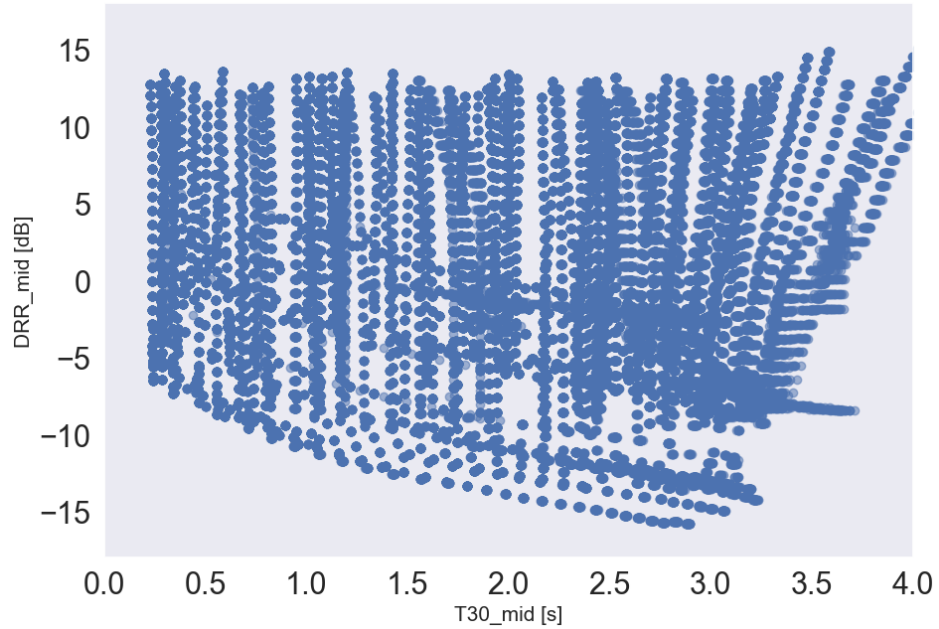

Figura 22. Gráfico de dispersión del $T30mid$ contra el DRR_{mid} para las RIR aumentadas.

En este último conjunto se observa finalmente un grupo de respuestas que simulan una variedad mucho más extensa de recintos comparados a los dos anteriores. Además, con este análisis, se valida concretamente el funcionamiento de los algoritmos para la aumentación tanto del T_{30} como del DRR.

Finalmente, en la tabla 3 se pueden observar los valores máximos y mínimos obtenidos para los descriptores T_{30} y DRR al juntar todos los conjuntos en una única base de datos. A su vez, en la Figura 23 se puede observar un gráfico de $T30_{mid}$ contra DRR_{mid} de toda la base de datos.

Tabla 3. Valores máximos y mínimos de los descriptores T_{30} y DRR mid.

	$T30_{mid}$ [s]	DRR_{mid} [dB]
Mínimo	0.18	-15.78
Máximo	4.09	16.77

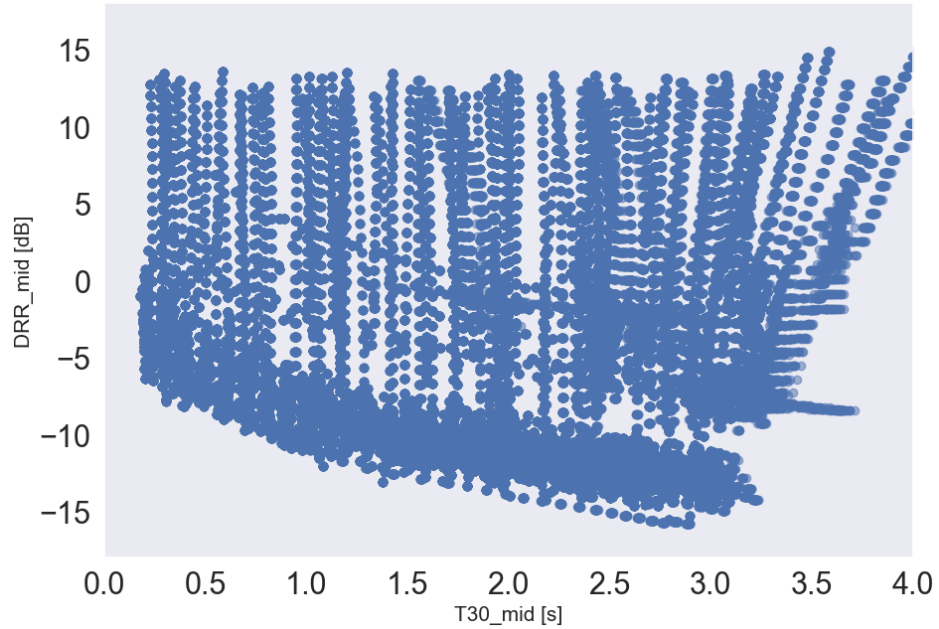

Figura 23. Gráfico de dispersión del $T30mid$ contra el DRR_{mid} para toda la base de datos.

Para tener un mejor entendimiento global de la base de datos, en las Figuras 24 y 25 se pueden observar unos diagramas de cajas de los valores de los descriptorees T_{30} y DRR para todas las bandas de frecuencias analizadas.

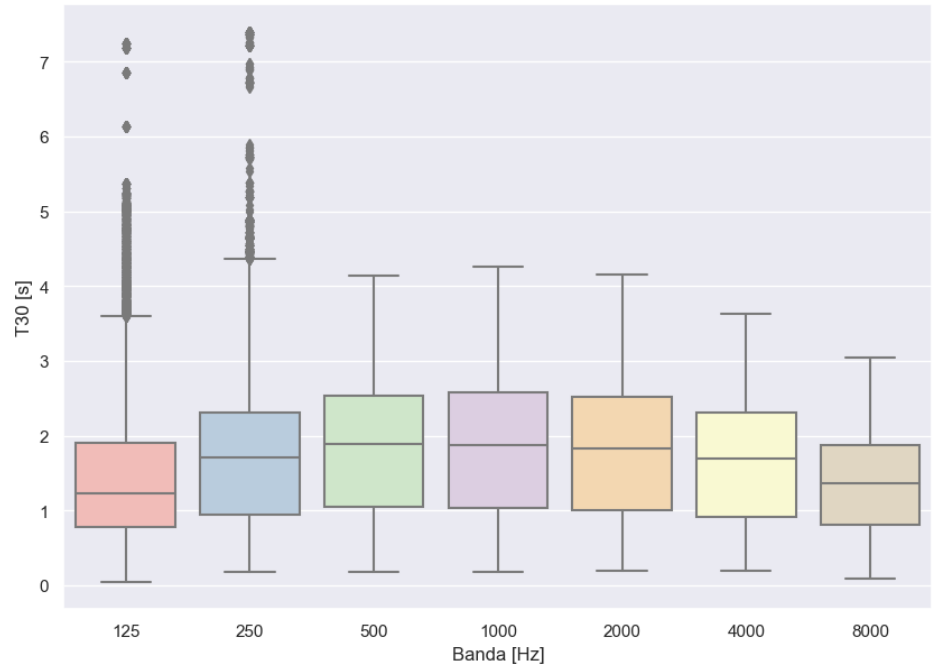

Figura 24. Boxplot de los valores de T_{30} por bandas de octava de toda la base de datos.

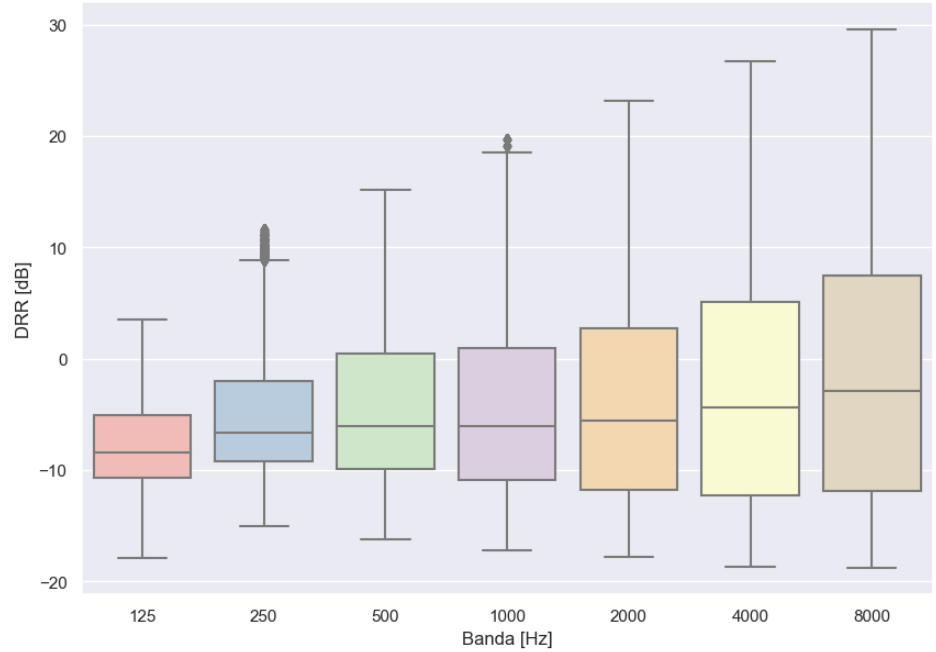

Figura 25. Boxplot de los valores de DRR por bandas de octava de toda la base de datos.

Se observaron una gran cantidad de valores atípicos en el cálculo del descriptor T_{30} de las respuestas al impulso para las bandas de 125 y 250 Hz. Estos son consecuencia del ruido intrínseco de las señales en esa banda en particular, el cual hace que el cálculo de la pendiente de cuadrados mínimos crezca y, por ende, el resultado final del parámetro también. No obstante, se optó por dejar estos valores en la base de datos para observar cómo se comporta la red frente a los mismos. En principio, analizando estos datos, se esperaría un mejor rendimiento del algoritmo en frecuencias altas.

Por otra parte, en la tabla 4 se pueden observar el total de respuestas al impulso que se conservaron en la base de datos luego del análisis y cálculo de los descriptores y, además, el total de horas de entrenamiento con las que se cuenta para cada banda.

Tabla 4. Cantidad de RIRs en la base de datos y total de horas de entrenamiento por bandas de frecuencia.

	Banda [Hz]						
	125	250	500	1000	2000	4000	8000
Total de RIRs [n]	3768	3929	4135	4141	4141	4145	4145
Horas de entrenamiento [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 conforme también aumenta la banda de frecuencia analizada. Nuevamente, viendo estos resultados, se espera un

mejor rendimiento del algoritmo en frecuencias altas debido a que cuenta con más horas de entrenamiento.

5.2. ENTRENAMIENTO DE LOS MODELOS

Una vez generadas y analizadas las bases de datos, se procede al entrenamiento de los modelos.

En concreto, se entrenaron 7 redes neuronales para la estimación por bandas de los parámetros acústicos a partir de señales de habla reverberadas sin ruido y otras 7 para la estimación utilizando señales con ruido agregado. Todas estas poseen la misma arquitectura, variando únicamente los datos de entrada y salida de las mismas en cada caso.

Para cada una de estas, se utilizaron un total de 500 épocas durante la etapa de entrenamiento, con el error cuadrático medio como función de costo. Además, del total de la base de datos, se destinó un 80 % para el entrenamiento, del cual, a su vez, se reservó un 10 % para la etapa de validación.

Utilizando la arquitectura descripta en la tabla 2 y esta elección de épocas, el algoritmo obtuvo las suficientes iteraciones para poder reducir su función de costo pero sin llegar a memorizar los datos. Esto se puede observar en la Figura 26, en donde tanto la función de costo como la de validación disminuyen, sin que esta última empiece a crecer (lo cual sería un comportamiento típico de que el modelo está haciendo un sobreajuste o, mejor conocido como, overfitting).

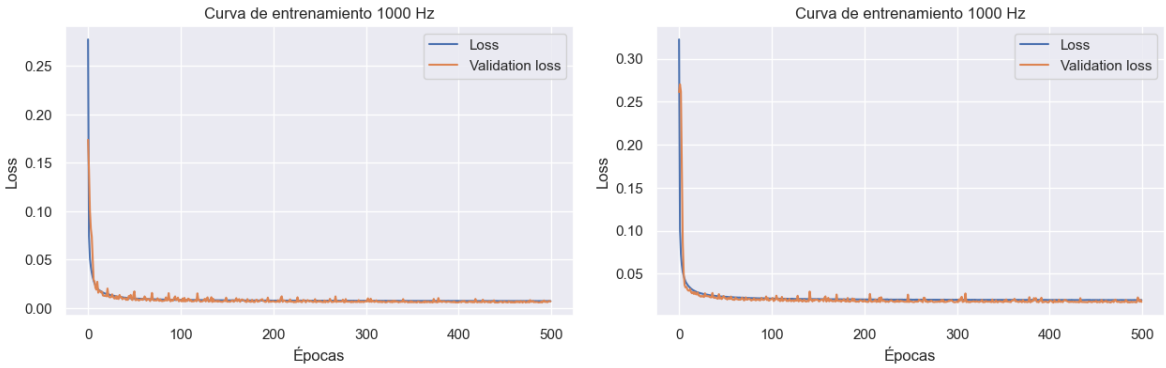

Figura 26. Curvas de entrenamiento para la banda de 1000 Hz de los modelos sin ruido y con ruido respectivamente.

Con esto, al menos en la parte de entrenamiento, se observó que el modelo propuesto es capaz de estimar los descriptores acústicos a partir de la TAE de un audio de voz reverbe-

rado, ya sea que esta tenga ruido o no. Este mismo comportamiento se vio reflejado en las demás bandas de estudio.

Aunque en ambos casos los valores de loss se minimizan, para detectar si existen diferencias significativas en la convergencia de los errores cuadráticos medios obtenidos, es necesario estudiar cada valor en particular. Para esto, en la tabla 5 se pueden observar los valores de loss obtenidos para cada banda de frecuencia en los modelos entrenados con TAE sin ruido y con ruido agregado respectivamente.

Tabla 5. Valores de Loss y Validation loss 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	
	Loss	Validation loss	Loss	Validation loss
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

Tal como se podía intuir observando la Figura 24, la banda de frecuencia con el mayor error cuadrático medio es la de 125 Hz. Esto es debido, principalmente, a la cantidad de valores atípicos que presenta la base de datos en esta frecuencia y la menor cantidad de horas de entrenamiento con la que cuenta. A su vez, se obtuvieron valores de loss considerablemente mejores para la base de datos sin ruido comparados a la que tiene ruido rosa agregado. Este comportamiento era esperable debido a que al agregar esta señal se disminuye el rango dinámico del audio y, por ende, también la dinámica de las pendientes de decaimiento de su TAE. Este fenómeno tiene una analogía directa con lo que sucede cuando en una sala se calcula su RIR y el cálculo del tiempo de reverberación se altera debido a que en la misma hubo ruido de fondo presente durante la medición.

5.3. EVALUACIÓN DEL MODELO CON AUDIOS DESCONOCIDOS

En esta sección, se busca determinar la eficiencia de los modelos para predecir los descriptores de la sala, probándolos con audios que no se utilizaron durante la etapa de entrenamiento.

En concreto, se calculó el coeficiente de correlación de Spearman (y su valor p asociado) tomando los valores que pudo predecir la red en contraste con los calculados a partir de las RIRs.

5.3.1. ANÁLISIS T_{30}

El primer descriptor elegido para realizar la comparación es el tiempo de reverberación, ya que se considera como uno de los más importantes para el análisis acústico de los recintos.

En la tabla 6 se pueden observar los coeficientes de correlación de Spearman entre las predicciones y los valores calculados directamente con las RIRs y los métodos normados.

Tabla 6. 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 sin ruido		Base de datos con ruido	
	r	Valor p	r	Valor p
125	0.716	<0.01	0.638	<0.01
250	0.937	<0.01	0.765	<0.01
500	0.954	<0.01	0.722	<0.01
1000	0.961	<0.01	0.813	<0.01
2000	0.957	<0.01	0.854	<0.01
4000	0.948	<0.01	0.886	<0.01
8000	0.949	<0.01	0.910	<0.01

Como se intuyó en la sección anterior observando los diagramas de caja de las bases de datos de este descriptor, la banda de frecuencia con la menor correlación es la de 125 Hz. No obstante, salvo por esta banda, en el caso de la base de datos sin ruido agregado todos los coeficientes están por encima del 90 %, lo cual demuestra que la red es capaz de estimar este descriptor con gran precisión.

Por otra parte, como era de esperar, los coeficientes de correlación utilizando la base de dato con ruido rosa agregado resultaron inferiores con respecto a los del otro conjunto de datos. No obstante, todos están por encima del 63 %, con lo cual se puede concluir que la red es capaz de estimar los descriptores aunque tengan ruido de fondo, aunque con menor precisión.

5.3.2. ANÁLISIS C_{50}

En segunda instancia, se analizan los valores de correlación obtenidos para el descriptor C_{50}. Dichos resultados se pueden observar en la tabla 7.

Tabla 7. 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 sin ruido		Base de datos con ruido	
	r	Valor p	r	Valor p
125	0.644	<0.01	0.669	<0.01
250	0.674	<0.01	0.665	<0.01
500	0.837	<0.01	0.762	<0.01
1000	0.912	<0.01	0.838	<0.01
2000	0.934	<0.01	0.839	<0.01
4000	0.970	<0.01	0.844	<0.01
8000	0.970	<0.01	0.844	<0.01

Al igual que en el caso del tiempo de reverberación, este descriptor obtuvo mejores resultados con la red entrenada con datos sin ruido agregado y, además, nuevamente se obtuvieron los peores resultados para la banda de 125 Hz.

Si bien solamente se realizó un análisis de valores atípicos encontrados en la banda de 125 Hz para el tiempo de reverberación, estos resultados dejan ver que existe una cierta correlación entre los elementos que afectan al cálculo de los descriptores. Es decir, la falta de rango dinámico en los audios también afecta a la estimación del descriptor C_{50}, lo cual no es un comportamiento que se vea de forma intuitiva al analizar la fórmula para su cálculo.

Como se ve en las siguientes dos secciones, este mismo fenómeno también se observa en los dos descriptores restantes. Esto puede deberse al hecho de que existe una correlación entre el cálculo de estos cuatro descriptores, la cual estaría dada porque todos provienen de la curva de decaimiento de la RIR.

5.3.3. ANÁLISIS C_{80}

En la tabla 8 se pueden observar los valores de los coeficientes de correlación de Spearman entre las predicciones del descriptor C_{80} y su cálculo convencional.

Tabla 8. 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 sin ruido		Base de datos con ruido	
	r	Valor p	r	Valor p
125	0.817	<0.01	0.768	<0.01
250	0.874	<0.01	0.790	<0.01
500	0.917	<0.01	0.816	<0.01
1000	0.952	<0.01	0.876	<0.01
2000	0.963	<0.01	0.881	<0.01
4000	0.984	<0.01	0.885	<0.01
8000	0.982	<0.01	0.877	<0.01

Al igual que en los dos descriptores anteriores, la red entrenada con la base de datos sin ruido obtuvo los mejores resultados de estimación, con todas las bandas por encima del 81%. A su vez, nuevamente, la banda con los peores resultados fue la de 125 Hz.

5.3.4. ANÁLISIS D_{50}

Por último, en la tabla 9 se observan los valores de los coeficientes de correlación de Spearman entre las predicciones y los valores calculados del descriptor D_{50}.

Tabla 9. 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 sin ruido		Base de datos con ruido	
	r	Valor p	r	Valor p
125	0.640	<0.01	0.666	<0.01
250	0.676	<0.01	0.671	<0.01
500	0.835	<0.01	0.764	<0.01
1000	0.905	<0.01	0.839	<0.01
2000	0.935	<0.01	0.839	<0.01
4000	0.967	<0.01	0.848	<0.01
8000	0.968	<0.01	0.839	<0.01

Siguiendo la tendencia de los descriptores anteriormente analizados, los valores con mejor estimación fueron los obtenidos con la red entrenada con la base de datos sin ruido rosa agregado. Y, como en los demás, el coeficiente de correlación más bajo se encontró en la banda de 125 Hz.

Siendo que los descriptores se calcularon a partir de audios de voz, cuyo rango de frecuencias se encuentra entre los 100 y 300 Hz aproximadamente, se intuía que las mayores correlaciones se den en estas bandas de baja frecuencia, pero esto no fue así para ningún

parámetro acústico. No es posible tener certezas de por qué sucede esto, pero se puede concluir que es porque el ruido intrínseco del audio es menor en altas frecuencias y por esta razón el rango dinámico crece, teniendo un mejor rango de decaimiento de la curva para usarlo en la estimación.

5.4. EVALUACIÓN DEL MODELO CON MEDICIONES REALES

Como último paso, se decidió evaluar los modelos con audios grabados en distintas salas reales. En concreto, se grabaron barridos frecuenciales y audios de voz en 3 salas distintas y en 2 puntos de medición en cada una.

De los barridos frecuenciales se calcularon las RIRs de las salas, de las cuales se obtuvieron los valores de los descriptores de las mismas con el software REW [39]. En paralelo, los audios de voz se procesaron para obtener los TAEs y usarlos para estimar los valores con los modelos entrenados.

Para asegurar que los audios de voz y los barridos frecuenciales se tomen desde la misma posición, ambos se emitieron utilizando el mismo parlante dentro de la sala.

Por último, se utilizaron los valores de JND (diferencia apenas perceptible, del inglés *just noticeable difference*) de los descriptores como parámetro para contrastar qué tan lejos están las estimaciones de una variación apenas perceptible. No obstante, al ser este un parámetro tan restrictivo, no se lo consideró como un valor determinante para decidir si la estimación fue buena o no. Los mismos se pueden encontrar en la tabla 10.

Tabla 10. JND de los descriptores acústicos.

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

5.4.1. SALA 1

Para la primera sala elegida, se grabaron barridos frecuenciales y audios de voz en dos puntos de la misma, promediando luego los resultados obtenidos.

En la tabla 11 se pueden observar los resultados estimados con la red neuronal en contraste con los valores calculados con métodos convencionales a partir de la respuesta al

impulso en la posición 1.

Tabla 11. 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

En paralelo, la tabla 12 muestra la misma información pero para la posición 2.

Tabla 12. 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

Para obtener un mejor entendimiento de las tablas anteriores, se decidió expresar las diferencias entre los valores estimados con ambas redes neuronales contra los descriptores obtenidos mediante los métodos convencionales. Estos resultados se expresan en la tabla 13.

Tabla 13. 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

Para esta primera sala, se obtuvieron valores estimados muy cercanos a los esperados, incluso muchos de ellos por dentro del límite de los JND. Ambas posiciones mostraron mediciones consistentes, salvo en el caso del descriptor D_{50} que fue en el cual se encontraron las dispersiones más significativas.

5.4.2. SALA 2

Tal cual como se realizó en la primera sala, en esta se grabaron barridos frecuenciales y audios de voz en dos puntos distintos, comparando luego los resultados. En la tabla 14 se pueden observar los valores estimados con la red neuronal y los valores obtenidos mediante las respuestas al impulso del recinto para la posición 1.

Tabla 14. 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

Por otra parte, en la tabla 15 se exponen los valores estimados por la red neuronal y los valores calculados a partir de las respuestas al impulso para la segunda posición.

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 2.

Obtenido de	Banda [Hz]	T_{30} [s]	C_{50} [dB]	C_{80} [dB]	D_{50} [%]
RIR		1.05	1.12	6.05	56.37
Audio de voz	125	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		1.31	0.52	4.70	50.50
Audio de voz	250	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		2.10	-4.50	-1.88	31.38
Audio de voz	500	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		1.93	-2.37	0.84	39.96
Audio de voz	1000	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		1.72	-2.80	0.35	38.84
Audio de voz	2000	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		1.21	-0.43	2.76	47.08
Audio de voz	4000	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		0.81	1.69	5.91	60.17
Audio de voz	8000	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

Finalmente, para poder comparar los resultados de las tablas anteriores, en la tabla 16 se pueden observar las diferencias entre los valores estimados en ambas posiciones con la red neuronal contra los obtenidos a partir de las respuestas al impulso del recinto.

Tabla 16. 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

En línea a lo visto en la sala anterior, se encontraron valores estimados muy cercanos a los calculados mediante métodos convencionales. Nuevamente, muchos de los resultados obtenidos están dentro del límite de los JND y el descriptor con las mayores dispersiones vuelve a ser el D_{50}.

5.4.3. SALA 3

Por último, al igual que en las dos salas anteriores, se grabaron audios de voz y se obtuvieron las respuestas al impulso de este recinto. Los resultados estimados con la red neuronal para la primera posición se pueden observar en la tabla 17.

Tabla 17. 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 ± 1.00	3.31 ± 1.00	44.96 ± 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 ± 1.00	3.56 ± 1.00	45.65 ± 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 ± 1.00	-1.67 ± 1.00	27.92 ± 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 ± 1.00	0.60 ± 1.00	38.03 ± 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 ± 1.00	0.74 ± 1.00	38.84 ± 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 ± 1.00	3.62 ± 1.00	49.35 ± 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 ± 1.00	6.12 ± 1.00	70.59 ± 5.00

A su vez, los valores obtenidos para la segunda posición se encuentran en la tabla 18.

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 2.

Obtenido de	Banda [Hz]	T_{30} [s]	C_{50} [dB]	C_{80} [dB]	D_{50} [%]
RIR		0.91	-0.78	2.99	47.77
Audio de voz	125	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		1.10	-1.84	2.09	40.12
Audio de voz	250	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		1.58	-4.07	-0.82	30.99
Audio de voz	500	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		1.49	-2.83	0.19	36.39
Audio de voz	1000	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		1.32	-2.48	0.28	36.21
Audio de voz	2000	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		0.98	-0.13	3.36	48.66
Audio de voz	4000	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		1.13	1.78	4.51	59.67
Audio de voz	8000	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

Por último, a modo de comparación, se contrastan los resultados de las tablas anteriores obteniendo las diferencias entre los valores estimados en ambas posiciones contra su referencia. Estos valores se pueden observar en la tabla 19.

Tabla 19. 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

Nuevamente, al igual que en los casos anteriores, los resultados con mayor dispersión entre lo estimado y lo esperado se encontraron en el descriptor D_{50}. No obstante, en esta sala también se obtuvieron muy buenos valores en ambas posiciones.

Una vez analizadas todos los recintos, al ver las tablas de diferencias y contrastarlas con los respectivos JND de cada descriptor, es posible determinar que la red neuronal es capaz de estimar los descriptores de las salas con mucha precisión. Aunque todos los recintos tienen dimensiones diferentes, no se observaron mejores estimaciones en ninguna de las salas, con lo cual se concluye que el algoritmo funciona de igual manera en diversos volúmenes de recintos, siempre y cuando los descriptores estén dentro de los valores con los que fue entrenado.

No se pudo determinar por qué el descriptor D_{50} tuvo mayores dispersiones que los otros tres parámetros, pero se estima que es debido a que es el único que no contempla una escala de decibeles en su cálculo.

Queda como trabajo futuro realizar más mediciones en recintos reales, pero esta vez controlando la relación señal/ruido de la sala para poner a prueba la red neuronal entrenada con ruido rosa. Con esto, sería posible determinar si la estimación de la red que se entrenó con estas condiciones sea más precisa en estos casos que la que no lo usó; esperando encontrar que esto sea afirmativo.

6. CONCLUSIONES

En esta investigación se implementó un algoritmo que utiliza procesamientos de señales de habla y redes neuronales convolucionales para la obtención ciega de parámetros acústicos de una sala. Las mismas se entrenaron con una base de datos de audios reverberados de forma artificial con respuestas al impulso de distintos recintos, teniendo como finalidad estimar los descriptores a partir de las envolventes de estas señales. A su vez, a estos mismos archivos se les agregó ruido rosa de forma tal de conseguir distintas relaciones señal/ruido controladas, con los cuales se entrenó la misma arquitectura de red con el fin de estudiar si esto podía mejorar las estimaciones en situaciones reales de medición.

A diferencia de otros algoritmos analizados en el estado del arte, el que se emplea en este trabajo tiene la posibilidad de estimar varios descriptores acústicos a la vez. Esta característica lo diferencia ampliamente de los demás que, en su mayoría, implementan el mismo algoritmo reiteradamente pero en cada caso lo entrenan para obtener un solo descriptor en concreto.

Dada la dificultad de encontrar una base de datos que represente todos los recintos posibles, se decidió estudiar la utilización de técnicas de generación de respuestas al impulso sintéticas y de aumentación de respuestas reales para lograr abarcar la mayor cantidad de salas que se puedan encontrar en situaciones de mediciones de campo reales.

En concreto, a lo largo de esta investigación y con los resultados obtenidos, se demostró que es posible implementar un algoritmo que sea capaz de estimar varios parámetros acústicos a la vez. Esto es posible ya que todos provienen de las curvas de decaimiento de los audios utilizados. Por otra parte, se validaron las técnicas utilizadas para la aumentación de respuestas al impulso reales y la generación de las sintéticas, comprobando además que el uso de las mismas generó amplias mejoras durante la etapa de entrenamiento debido a que permiten agrandar y homogeneizar la base de datos.

Al utilizar los datos de validación, se encontró que la red neuronal entrenada con datos sin agregarles ruido rosa fue la que obtuvo mejores resultados de estimación en todas las bandas de frecuencia. En cuanto a la prueba con mediciones de campo, se obtuvieron buenas estimaciones de los parámetros, incluso muchas de ellas están dentro del rango de los JND. No obstante, el descriptor D_{50} fue el que obtuvo las mayores dispersiones, con lo

cual queda en evidencia que todavía queda bastante por mejorar para ajustar el algoritmo.

En conclusión, los resultados de esta tesis demuestran que aún queda mucho por estudiar en el área de la estimación ciega de parámetros acústicos. El haber comprobado que es posible estimar varios descriptores a la vez abre la posibilidad de la creación y mejora de muchos algoritmos ya existentes, de los cuales se espera que este trabajo sirva como inspiración.

7. TRABAJO FUTURO

Los resultados obtenidos durante esta investigación se consideran muy satisfactorios ya que se pudo demostrar que es posible estimar múltiples parámetros acústicos a la vez con una única red neuronal. No obstante, hay cierto lugar a mejoras para implementar en el proceso.

En principio, las técnicas de aumentación de respuestas al impulso utilizadas solo permiten variar el tiempo de reverberación o la relación directo-reverberado. Con esto se genera un control preciso del TR en la base de datos, pero dejan a la claridad y la definición con valores aleatorios. Es por esta razón que la red tiene mayor porcentaje de aciertos en la estimación del T_{30} y menos en el C_{50}, C_{80} y D_{50}. Por ende, sería conveniente implementar técnicas de aumentación que permitan la manipulación de estos parámetros también.

Por otro lado, no se realizó un estudio profundo en cuanto a la arquitectura de red elegida. Por tanto, es posible que una variación de la misma, o la implementación de otro tipo completamente distinto, permita obtener mejores estimaciones.

Por último, las mediciones de campo que se hicieron fueron realizadas en condiciones favorables según lo establecido en la norma. Es decir, sin mucha presencia de ruido en el lugar. En cierto sentido esto pondera a que la red neuronal que se entrenó sin agregarle ruido a los datos pueda llegar a tener mejores resultados en la estimación, cosa que se observó en el análisis de esta investigación. Por esto, sería conveniente agregar fuentes externas y controladas de ruido en futuras mediciones para determinar si esta es verdaderamente mejor o si el hecho de haber usado ruido rosa durante el entrenamiento genera que se puedan lograr mejores estimaciones en condiciones adversas.

BIBLIOGRAFÍA

[1] V. G. Escobar y J. B. Morillas, “Analysis of intelligibility and reverberation time recommendationsin educational rooms,” *Applied Acoustics 96*, págs. 1-10, (2015).

[2] W. J. Murphy y N. Xiang, “Room acoustic modeling and auralization at an indoor firing range,” *The Journal of the Acoustical Society of America*, págs. 3868-3872, (2019).

[3] M. Țopa, N. Toma, B. Kirei e I. Crişan, “Evaluation of Acoustic Parameters in a Room,” *WSEAS International Conference on SIGNAL PROCESSING*, págs. 41-44, (2010).

[4] “ISO 3382-1:2009, Acoustics. Measurement of room acoustic parameters. Part 1: Performance spaces,” International Organization for Standardization, Standard, jun. de (2009).

[5] “IEC 60268-16:2020, Sound system equipment. Part 16: Objective rating of speech intelligibility by speech transmission index,” International Electrotechnical Commission, Standard, sep. de (2021).

[6] A. V. Oppenheim y A. S. Ian, *Signals and systems*, 2.ª ed. Prentice-Hall, Inc, (1996), págs. 74-137.

[7] N. M. Papadakis y G. E. Stavroulakis, “Review of Acoustic Sources Alternatives to a Dodecahedron Speaker,” *MDPI*, págs. 1-32, (2019).

[8] J. Pätynen, B. F. Katz y T. Lokki, “Investigations on the balloon as an impulse source,” *The Journal of the Acoustical Society of America*, vol. 129, n.º 1, EL27-EL33, (2011).

[9] S. Duangpummet, J. Karnjana, W. Kongprawechnon y M. Unoki, “Blind Estimation of Room Acoustic Parameters and Speech Transmission Index using MTF-based CNNs,” *ELSEVIER*, págs. 1-12, (2021).

[10] P. Kendrick, F. F. Li y T. J. Cox, “Blind Estimation of Room Acoustic Parameters and Speech Transmission Index using MTF-based CNNs,” *ACTA Acustica United With Acustica*, págs. 1-11, (2007).

[11] ——, “Monaural room acoustic parameters from music and speech,” *The Journal of the Acoustical Society of America*, págs. 1-11, (2008).

[12] *THE ACE CHALLENGE*, http://www.ee.ic.ac.uk/naylor/ACEweb/index.html, Extraído el 18 de abril del 2022.

[13] J. Eaton, N. D. Gaubitch, A. H. Moore y P. A. Naylor, "Estimation of room acoustic parameters: The ACE Challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, págs. 1-10, (2016).

[14] P. P. Parada, D. Sharma, T. van Waterschoot y P. A. Naylor, "Evaluating the Non-Intrusive Room Acoustics Algorithm with the ACE Challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, págs. 1-5, (2015).

[15] T. de M. Prego, A. A. de Lima, R. Zambrano-López y S. L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, págs. 1-5, (2015).

[16] H. Loellmann, A. Brendel, P. Vary y W. Kellermann, "Single-Channel Maximum-Likelihood T60 Estimation Exploiting Subband Information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, págs. 1-5, (2015).

[17] N. J. Bryan, "Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation," *Adobe Research*, págs. 1-5, (2019).

[18] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," (2000).

[19] S. W. Smith, *The scientist and engineer's guide to digital signal processing*, 6.ª ed. Calif: California Technical Pub, (1997), págs. 277-284.

[20] Y.-W. Liu, *Fourier Transform Applications*. InTech, (2012), págs. 291-300.

[21] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, n.º 6, págs. 1187-1188, (1965).

[22] A. Lundeby, T. E. Vigran, H. Bietz y M. Vorländer, "Uncertainties of measurements in room acoustics," *Acta Acustica united with Acustica*, vol. 81, n.º 4, págs. 344-355, (1995).

[23] M. R. Schroeder, "Modulation transfer functions: Definition and measurement," *Acta Acustica united with Acustica*, vol. 49, n.º 3, págs. 179-182, (1981).

[24] A. S. 1-2013, *Acoustical Terminology*, (2013).

[25] I. A. Basheer y M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *Journal of microbiological methods*, vol. 43, n.º 1, págs. 3-31, (2000).

[26] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, n.º 6, pág. 386, (1958).

[27] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural networks*, vol. 4, n.º 2, págs. 251-257, (1991).

[28] T. Nishijima, "Universal approximation theorem for neural networks," *arXiv preprint arXiv:2102.10993*, (2021).

[29] F. Chollet, *Deep learning with Python*. Simon y Schuster, (2021).

[30] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. .º'Reilly Media, Inc.", (2019), págs. 431-461.

[31] *Delft University of Technology*, https://www.tudelft.nl/en/, Extraído el 2 de Febrero del 2023.

[32] R. Stewart y M. Sandler, "Database of omnidirectional and B-format room impulse responses," págs. 165-168, (2010).

[33] D. P. Kingma y J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, (2014).

[34] Martín Abadi, Ashish Agarwal, Paul Barham y col., *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, https://www.tensorflow.org/, Extraído el 27 de Enero del 2023.

[35] L. Myers y M. J. Sirois, "Spearman correlation coefficients, differences between," *Encyclopedia of statistical sciences*, vol. 12, (2004).

[36] *Behringer ECM8000*, https://www.behringer.com/product.html?modelCode=P0118, Extraído el 18 de Enero del 2023.

[37] *Sony MHC-EC99*, https://www.sony.es/electronics/support/audio-systems-mhc-series/mhc-ec99, Extraído el 18 de Enero del 2023.

[38] M. A. Ortiz, *Repositorio de trabajo: .Estimación ciega de parámetros acústicos de un recinto"*, https://github.com/maxiaortiz22/blind-estimation-of-acoustics-parameters, Extraído el 24 de Noviembre del 2022.

[39] *REW: Room Acoustics Software*, https://www.roomeqwizard.com/, Extraído el 10 de Enero del 2023.