UNTREF UNIVERSIDAD NACIONAL DE TRES DE FEBRERO

INGENIERÍA DE SONIDO

Estimación ciega de parámetros acústicos de un recinto

Tesis final presentada para obtener el título de Ingeniero de Sonido de la Universidad Nacional de Tres de Febrero (UNTREF)

TESISTA: Maximiliano Adriel Ortiz (40.440.819)
TUTOR/A: Nombre y apellido
COTUTOR/A: Martin Bernardo Meza

Fecha de defensa: mes y año | Locación (ej. Sáenz Peña), Argentina

AGRADECIMIENTOS

Se propone incluir este apartado, donde se debe agradecer primeramente a las autoridades de la Universidad, al coordinador de la carrera, al tutor y a los docentes implicados en el desarrollo de la investigación. Seguidamente agradecer a familiares o a aquellas personas que se quiera. También puede incluirse en la siguiente hoja una dedicatoria personal. A modo de ejemplo el contenido podría ser:

"En primer lugar dar gracias a la Universidad Nacional de Tres de Febrero (UNTREF), a su Rector Lic. Anibal Jozami, a todo su personal docente y no docente. Por promover un espacio ideal para el desarrollo de ideas y nuevos pensamientos y brindar a todos y cada uno de los alumnos, de esta casa de altos estudios, todos los recursos que esta institución dispone.

Esta investigación no hubiera sido posible sin una formación académica acorde, por este motivo debo extender mi agradecimiento a los docentes de la carrera de Ingeniería de Sonido de la UNTREF, a su coordinador Ing. Alejandro Bibondo, que siendo la primera carrera de estas características del país, es muy importante contar con un cuerpo docente afín a las exigencias que este desafío propone, prestando su dedicación y vocación de enseñar.

Un especial agradecimiento por la participación de esta tesis a la tutora Ing. Nombre Apellido, que supo transmitirme sus conocimientos y ayudarme a organizarme y fijarme un rumbo concreto y delineado, disponiendo desmedidamente de su tiempo.

Por otra parte, quisiera hacer una mención especial al Ing. Hernan San Martin, que permitió el uso de las instalaciones de su laboratorio para poder trabajar y la disposición de todos sus recursos para que dicha investigación se realizara en tiempo y forma.

Por último y no menos importante, quiero dar un afectuoso y cálido agradecimiento a mi familia…"

DEDICATORIA

Elige a quién o a qué quieres dedicárselo

Elegir el motivo de la dedicatoria (orientativo)

ÍNDICE DE CONTENIDOS

RESUMEN vi

ABSTRACT vii

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 DE LA SALA 9

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

3.2.2 CLARIDAD: C_{80} 12

3.2.3 DEFINICIÓN: D_{50} 12

3.2.4 TIEMPO CENTRAL: T_S 13

3.2.5 ÍNDICE DE TRANSMISIÓN DE VOZ: STI 13

3.3 RESPUESTAS AL IMPULSO SINTÉTICAS 15

3.4 CURVA DE DECAIMIENTO DE UNA SEÑAL DE VOZ 17

3.5 ENVOLVENTE TEMPORAL DE AMPLITUD: TAE 19

3.6 REDES NEURONALES ARTIFICIALES 20

3.6.1 EL PERCEPTRÓN MULTICAPA 20

3.6.2 ENTRENAMIENTO DE UNA RED NEURONAL 21

3.6.3 REDES NEURONALES CONVOLUCIONALES: CNN 22

4. METODOLOGÍA 25

4.1 GENERACIÓN DE BASE DE DATOS 25

4.2 MODELO PROPUESTO 26

4.3 EVALUACIÓN DEL MODELO 28

5. RESULTADOS Y ANÁLISIS 29

5.1 ENTRENAMIENTO DE LOS MODELOS 29

5.2 EVALUACIÓN DEL MODELO CON AUDIOS DESCONOCIDOS 30

6. CONCLUSIONES 33

7. LÍNEAS FUTURAS DE INVESTIGACIÓN 34
BIBLIOGRAFÍA 35

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. 12
Figura 3. Diagrama en bloques para el cálculo del STI. 15
Figura 4. Diferencias en la curva de decaimiento para un audio limpio contra uno reverberado. 18
Figura 5. Pasos para la obtención de la TAE. 20
Figura 6. Diagrama en bloques del modelo propuesto. 26
Figura 7. Curvas de entrenamiento de la red de la banda de 1 kHz. 30
Figura 8. Rectas de regresión entre los valores predichos y los reales para la banda de 1 kHz……. 31

ÍNDICE DE TABLAS

Tabla 1. Factores de ponderación por banda de octava para los índices de transmisión de modulación 15
Tabla 2. Arquitectura de red propuesta. 27
Tabla 3. Coeficiente de correlación de Spearman entre las predicciones y el valor real de T30 por cada banda. 31

ÍNDICE DE ANEXO

A 1. Curvas de entrenamiento de los modelos: 1
A 2. Rectas de regresión de los resultados obtenidos: 3

ÍNDICE DE FIGURAS ANEXO

Figura 1. Curvas de entrenamiento de la red de la banda de 125 Hz 1
Figura 2. Curvas de entrenamiento de la red de la banda de 250 Hz 1
Figura 3. Curvas de entrenamiento de la red de la banda de 500 Hz 2
Figura 4. Curvas de entrenamiento de la red de la banda de 2 kHz 2
Figura 5. Curvas de entrenamiento de la red de la banda de 4 kHz 3
Figura 6. Curvas de entrenamiento de la red de la banda de 8 kHz 3
Figura 7. Rectas de regresión entre los valores predichos y los reales para la banda de 125 Hz. 4
Figura 8. Rectas de regresión entre los valores predichos y los reales para la banda de 250 Hz. 4
Figura 9. Rectas de regresión entre los valores predichos y los reales para la banda de 500 Hz. 5
Figura 10. Rectas de regresión entre los valores predichos y los reales para la banda de 2 kHz. 5
Figura 11. Rectas de regresión entre los valores predichos y los reales para la banda de 4 kHz. 6
Figura 12. Rectas de regresión entre los valores predichos y los reales para la banda de 8 kHz……. 6

ÍNDICE DE TABLAS ANEXO

No se encuentran elementos de tabla de ilustraciones.

RESUMEN

El objetivo de esta investigación es estudiar un método que permite la obtención de parámetros acústicos a partir de señales de voz grabadas en un recinto, evitando la medición de la respuesta al impulso de la sala. Para esto, se propone un método, basado en el estado de arte actual, que utiliza redes neuronales convolucionales para estimar los parámetros necesarios para construir la respuesta al impulso de la sala a partir del método de Shroeder expendadido. Para entrenar a la red, se ingresa con la información de la envolvente temporal de amplitud (ETA) de una señal de habla reverberada. Una vez estimada la respuesta al impulso, es posible calcular los parámetros acústicos a partir de sus fórmulas conocidas. En concreto, en esta investigación se busca estimar los descriptores: índice de transmisión de voz (STI: Speech Transmission Index), tiempo de decaimiento temprano (EDT: Early decay time), tiempo reverberación (T60 y T30), claridad (C80) y definición (D50). Para lograr entrenar el modelo 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 impulsos reales y otros generados artificialmente. Finalmente, para probar la viabilidad del modelo se comparan con los parámetros obtenidos en una sala mediante métodos convencionales en contraste con el propuesto.

Palabras Clave: tiempo de reverberación, índice de transmisión de voz, parámetros acústicos de una sala, función de transferencia modulada, envolvente temporal de amplitud.

ABSTRACT

The aim of this research is to study a method that allows obtaining acoustic parameters from voice signals recorded in a room, avoiding the measurement of the impulse response of the room. For this, based on the current state of the art, a method which uses convolutional neural networks is proposed to estimate the necessary parameters to build the impulse response of the room from the Shroeder's expended method. To train the network, it is use amplitude time envelope (ETA) of a reverberated speech signal as input. Once the impulse response has been estimated, it is possible to calculate the acoustic parameters from their known formulas. Specifically, this research seeks to estimate the descriptors: speech transmission index (STI), early decay time (EDT), reverberation time (T60 and T30), clarity (C80) and definition (D50). In order to train the proposed model, a reverberated voice audio database is generated from the convolution of anechoic voice signals with a bank of real and artificially impulses generated. Finally, to test the feasibility of the model, they are compared with the parameters obtained in a room using conventional methods in contrast to the one proposed.

Keywords: reverberation time, speech transmission index, room acoustic parameters, modulated transfer function, temporal amplitude envelope

1. INTRODUCCIÓN

1.1 FUNDAMENTACIÓN

Para los ingenieros de sonido, la medición de parámetros acústicos de una sala es fundamental para entender objetivamente el comportamiento de esta 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 para una sala de reuniones o un recinto que permita entender perfectamente las señales sonoras de emergencia [1-2], los parámetros acústicos son la mejor herramienta con la que cuentan los ingenieros de audio para lograr llegar a esto [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 (C80), entre otras [4-5]. Los parámetros anteriormente mencionados (y la mayoría de los demás no descritos), 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 de este frente a un estímulo conocido [6].

Existen diversos métodos para obtener la respuesta al impulso de una sala. La idea tras su obtención es generar una señal de corta duración y gran amplitud, que sea capaz de excitar la sala en todas las frecuencias. Entre los más conocidos están: disparo en blanco, explosión de globos, aplausos, choque con maderas, barrido frecuencial, entre otros [7]. Sea cual sea el 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 esta razón, en esta investigación se propone un método de medición de parámetros acústicos que consiste en la estimación de estos 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 las 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 [8].

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 calcular los parámetros acústicos de una sala a partir de una 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 y modelados acústicos 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.
- 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 tres temáticas principales: la función de transferencia de modulación (MTF), los parámetros acústicos que se van a estimar (a saber: T30, DRR, Ts, C80 y D50) 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, T10, T20, T30, C80, D50 [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, sigue siendo muy costoso y complejo conseguirlas. Esto se debe a lo complejo de conseguir los equipos para su medición y, sobre todo, a que la medición se ve fuertemente afectada 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 el cálculo 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. [9], en la cual los autores presentaron un método para calcular el tiempo de reverberación de una sala a partir de un audio grabado en el lugar. Al no calcularse a partir de la RIR, los autores denominaron este método como una estimación ciega. Para la norma ISO 3382 [4], este descriptor se calcula 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 T10, -5 a -25 para el T20 y -5 a -35 para el T30). Tomando esta lógica, los autores descubrieron que existe cierta similitud 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 descrito 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 T20, EDT, C80 y Ts por bandas de frecuencia [10]. 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 aproximarse) 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* creó un desafío llamado *Acoustic Characterisation of Environments Challenge* (ACE Challenge) [11] en el año 2015. El objetivo de este desafío era evaluar algoritmos de última generación para la estimación ciega del tiempo de reverberación (Tr) 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 tomados en esas mismas salas [12]. Sumado a esto, la base de datos también cuenta con ruidos de ventiladores, balbuceos y ruido ambiente para juntar con los audios de voz y así poder determinar qué tan bien funcionan los algoritmos cuando las grabaciones tienen ruido.

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. [13] que utilizaron redes neuronales recurrentes para ir estimando tanto el Tr como el DRR en pequeños sectores del mismo audio aprovechando la propiedad de memoria de este tipo de redes, el de Prego et. al [14] 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 Tr y DRR encontrados (método que se podría considerar como una expansión del propuesto por Kendrick et. al. [9] y por último el propuesto por Loellman et. al. [15] que utiliza un MLE (inspirado, al igual que el trabajo anterior, en el trabajo de Kendrick et. al. [9]) para estimar los valores de Tr y DRR.

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 audios de salas con características diferentes a las que fueron entrenados. 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 costosa, Bryan propuso un método de aumentación de respuestas al impulso con la finalidad de obtener la mayor cantidad de RIRs posibles y así poder modelar muchas salas con el menor esfuerzo posible [16]. 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 Tr y otra red distinta cuyo resultado era el DRR. No fue hasta el 2021 que Duangpummet et. al. [8] 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 Tr 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.

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 de los sistemas es muy útil para muchas áreas dentro de la ingeniería y la acústica no es la excepción.

Si definimos un sistema 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.

En este caso, como la captación del sistema se hace a partir de un micrófono de medición, la excitación del sistema 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 (LSS)

De todos los métodos anteriormente mencionados, el método 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 controlar las características del estímulo como: duración, rango de frecuencias y amplitud.

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).

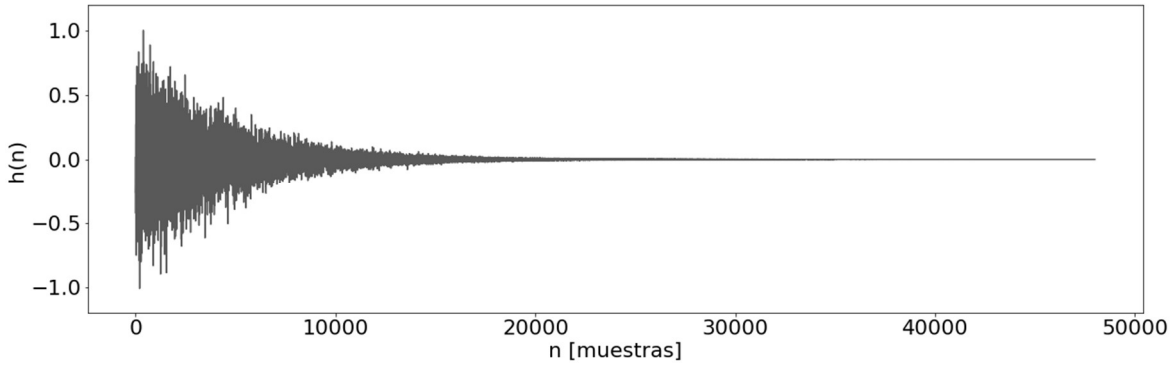

Figura 1. Ejemplo de respuesta al impulso de una sala.

Entendiendo a la sala como un sistema y conociendo su respuesta al impulso, es posible calcular la respuesta frente a un impulso generado en la misma y captado por un micrófono en su interior. En concreto, los estímulos medidos dentro del recinto se pueden representar como 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 una señal de prueba 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 grabado en una sala anecoica, 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 DE LA 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, C_{80}, D_{50} y T_s, y de la IEC 60268 [50] para el STI.

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 T60.

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 T60 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.
- T10: El rango de la curva desde 5 dB y 15 dB por debajo del máximo.
- T20: El rango de la curva desde 5 dB y 25 dB por debajo del máximo.
- T30: 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 decibelica. 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 [17], la transformada de Hilbert [18] y el método de integración de Schroeder [19].

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 esta nunca será así ya que el ruido es intrínseco, y variable, en cada 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 [20] 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, T10, T20 y T30 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, T10, T20 o T30 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.

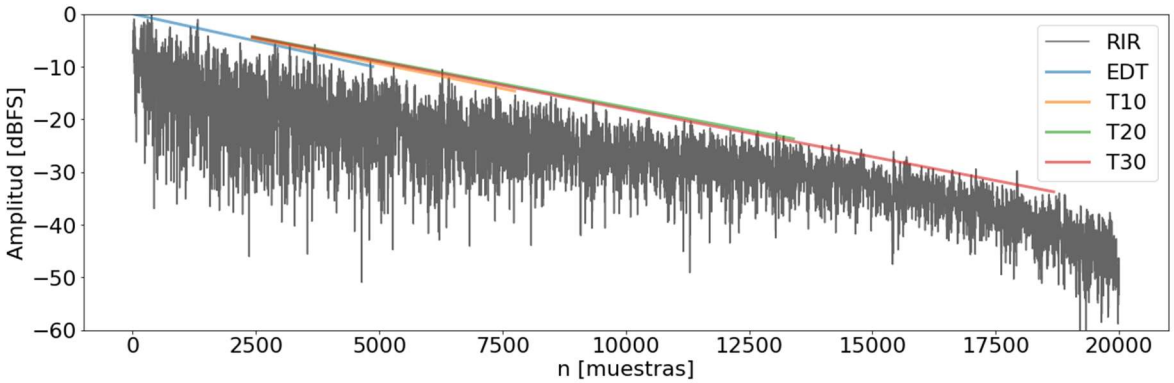

Figura 2. Decaimiento de una respuesta al impulso y sus aproximaciones por cuadrados mínimos para los descriptores EDT, T10, T20 y T30.

3.2.2 CLARIDAD: C_{80}

La claridad es uno de los descriptores más utilizados 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.

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

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.

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

3.2.4 TIEMPO CENTRAL: T_S

El descriptor de tiempo central (en inglés: center time) es el periodo en el centro de gravedad de una RIR. Este descriptor muestra el equilibrio entre claridad y reverberación relacionado con la inteligibilidad del habla.

Si se tiene una respuesta al impulso h(t), el descriptor T_s se puede calcular mediante la ecuación 6.

$$T_s = \frac{\int_0^\infty h^2(t)tdt}{\int_0^\infty h^2(t)dt} \tag{6}$$

3.2.5 ÍNDICE DE TRANSMISIÓN DE VOZ: STI

El índice de transmisión del habla (del inglés speech transmission index o STI) es un descriptor objetivo que se utiliza para predecir la inteligibilidad del habla y la dificultad de la escucha dentro de un recinto. La idea es caracterizar la calidad del canal de transmisión de un hablante a un oyente mediante un número de señal.

Este índice se calcula a través de la función de transferencia modulada (MTF) de un sistema [21-22]. Similar a una función de transferencia normal, la MTF representa las características de un canal de transmisión pero en función de una frecuencia de modulación y una disminución en la profundidad de modulación [23].

El concepto de las MTF es muy útil en la ingeniería acústica ya que sirve para cuantificar los efectos de la reverberación en una sala. En concreto, cuanto mayor sea la reverberación, menor resulta la profundidad de modulación de las señales moduladas que pasen a través del recinto. Para generalizar esto, se definen como índices de modulación a las relaciones de distorsión de modulación entre las envolventes de entrada y las salidas.

Dada una respuesta al impulso h(t), la magnitud de la MTF se calcula mediante la ecuación 7.

$$m(f_m) = \frac{\left|\int_0^\infty h^2(t) e^{-j2\pi t f_m} dt\right|}{\int_0^\infty h^2(t) dt} \quad (7)$$

donde m(f_m) es la MTF a la frecuencia de modulación f_m y h(t) es la respuesta al impulso de la sala.

Se utilizan un total de 98 estímulos modulados para calcular las relaciones de distorsión entre las señales emitidas y las grabadas en la medición. En concreto, los estímulos son señales moduladas en amplitud con siete bandas portadoras (los centros de octava de 125 a 8k Hz) y 14 frecuencias de modulación. Sabiendo esto, la relación de distorsión de modulación N se calcula a partir de la ecuación 8.

$$N_{k,i} = 10Log_{10}\left(\frac{m\left(f_{m_{i,k}}\right)}{1-m\left(f_{m_{i,k}}\right)}\right) \quad (8)$$

donde i es un número entero del 1 al 14 y representa las 14 frecuencias de modulación, k va del 1 al 7 y representa las 7 bandas portadoras.

El siguiente paso es determinar los índices de transmisión de modulación por banda, M_k. El mismo se calculó a partir del promedio de las relaciones de distorsión, como se observa en la ecuación 9.

$$M_k = \frac{1}{14}\sum_{i=1}^{14} N_{k,i} \quad (9)$$

Finalmente, el descriptor STI puede calcularse a partir de los valores M_k utilizando la ecuación 10.

$$STI = \sum_{k=1}^{7} \alpha_k M_k - \sum_{k=1}^{6} \beta_k \sqrt{M_k M_{k+1}} \quad (10)$$

donde α_k y β_k son los factores de ponderación para la banda de octava k, las cuales se pueden observar en la Tabla 1.

Tabla 1. Factores de ponderación por banda de octava para los índices de transmisión de modulación

Banda [Hz]	125	250	500	1000	2000	4000	8000
α	0.085	0.127	0.230	0.233	0.309	0.224	0.173
β	0.085	0.078	0.065	0.011	0.047	0.095	-

A modo ilustrativo, en la Figura 3 se pueden observar los pasos para el cálculo del STI.

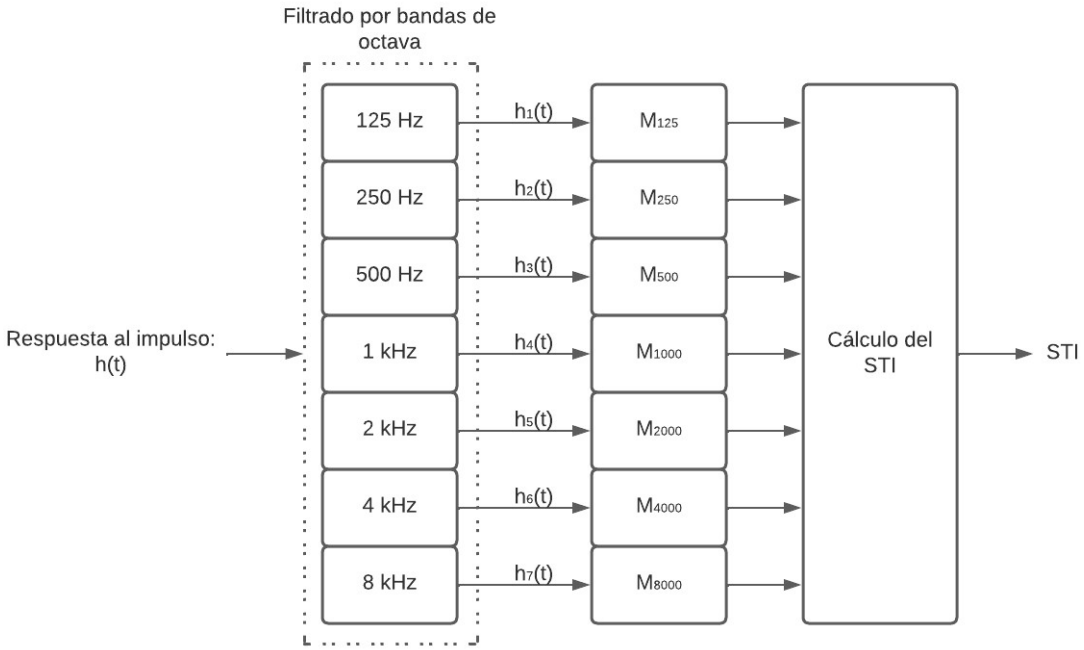

Figura 3. Diagrama en bloques para el cálculo del STI.

3.3 RESPUESTAS AL IMPULSO SINTÉTICAS

Si suponemos una sala con forma de paralelepípedo y sin obstáculos para la onda sonora en su interior, la forma de decaimiento en decibeles de su RIR va a ser de una pendiente perfecta. Esto quiere decir que el decaimiento del nivel se da de manera

homogénea y no se ve afectado por reflexiones tempranas. A este tipo de RIRs las denominaremos estocásticas.

Para modelar una RIR estocástica, es posible utilizar el método de Schroeder [24]. Este método supone que la señal puede ser representada por un ruido blanco gausiano cuya amplitud está dominada por una exponencial decreciente (notar que la exponencial en escala decibélica se transforma en una recta).

Por tanto, la respuesta al impulso por banda de frecuencia se puede modelar a través de la ecuación 11.

$$h(t) = ae^{\frac{-6.9}{T_{60}}} c_h(t) \tag{11}$$

donde a es un factor de amplitud, T_{60} es el tiempo de reverberación de la sala en una banda de frecuencia y $c_h(t)$ es un ruido blanco gaussiano filtrado con un pasa bandas de tercios de octava centrado en la banda principal de la RIR.

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 12.

$$h(t) = \sum_{k=1}^{n} e^{\frac{-6.9}{T_{60,k}}} c_{h,k}(t) \tag{12}$$

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 porque los campos de sonido en las habitaciones no son completamente difusos en la mayoría de los casos, se puedo contemplar un modelo de

síntesis que utilice el decaimiento de más de una exponencial para contemplar este fenómeno [9]. Al contemplar varias exponenciales a la vez, el resultado de la curva dará una cuyo decaimiento ya no sea exponencial.

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 gaussiano (r(t)), tal como se observa en la ecuación 13.

$$h(t) = e(t)r(t) \tag{13}$$

Ahora, la envolvente se representa como una suma de envolventes. Esto se puede observar en la ecuación 14.

$$e(t) = \sum_{k=1}^{M} \alpha_k a_k^n \tag{14}$$

donde a_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. Por tanto, la ecuación de la envolvente puede simplificarse como se observa en la ecuación 15.

$$e(t) = \sum_{k=1}^{M} \alpha_k a_k^n \tag{15}$$

3.4 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 de este.

Sabiendo esto, en la investigación de Kendrick et al. [9] 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 4 se compara la curva de decaimiento de una señal contra sí misma pero afectada por la reverberación de un recinto.

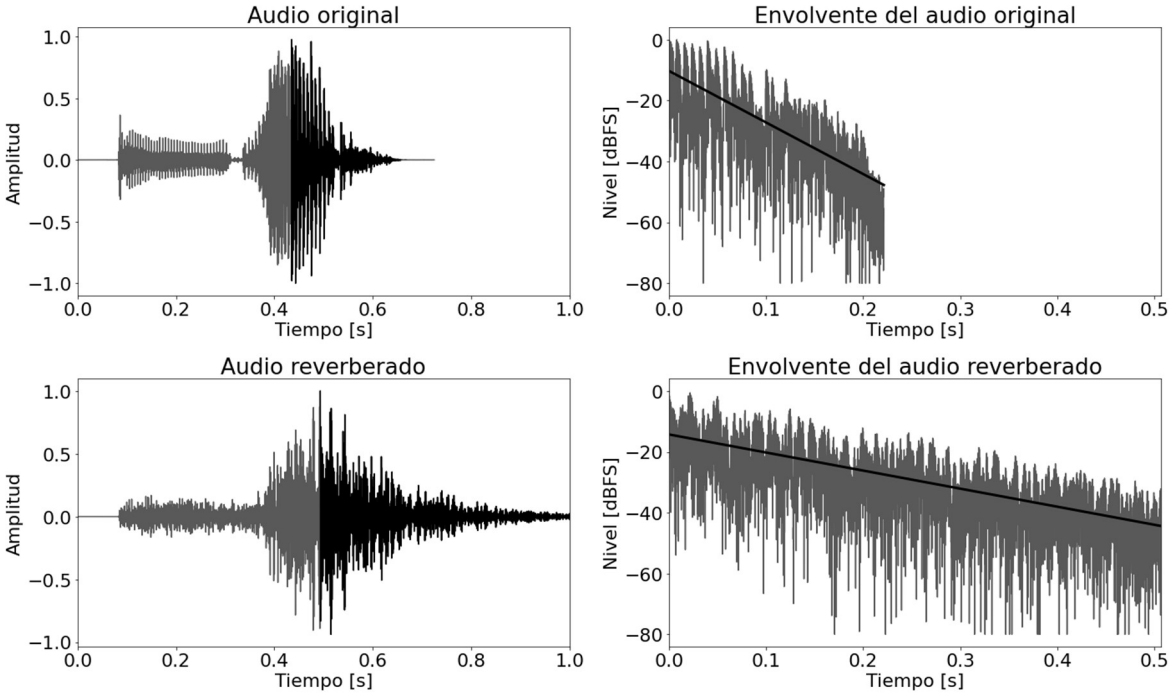

Figura 4. Diferencias en la curva de decaimiento para un audio limpio contra uno reverberado.

3.5 ENVOLVENTE TEMPORAL DE AMPLITUD: TAE

Según lo expuesto en la sección 3.4, se observó que el decaimiento de una señal de voz captada en un recinto guarda información de este.

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, se plantea encontrar 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 contenido se logran ciertas ventajas: en principio, podemos independizarnos del contenido del audio, ya que lo único importante en el estudio van a ser las partes de decaimiento de la señal de las que se pueden obtener los parámetros acústicos (tal como se vio en la sección anterior). En segundo lugar, la información se comprime drásticamente y se elimina todo ruido que pueda interferir en el entrenamiento de una red neuronal (más detalles sobre esto en la sección siguiente).

Para obtener la TAE se deben realizar los siguientes pasos:

- **Obtener una señal reverberada:** en esta investigación en particular se utilizan audios de 5 s de duración con una frecuencia de muestreo de 16 kHz.
- **Filtrar la señal con un filtro pasa banda:** en esta tesis se usaron los cetros de octava de 125, 250, 500, 1000, 2000, 4000 y 8000 Hz.
- {Obtener la transformada de Hilbert de la señal:} esta devuelve la envolvente de una señal.
- **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.
- **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 muestreo 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.

- **Normalizar la señal obtenida.**

Los pasos descriptos anteriormente se pueden observar de forma gráfica en la Figura 5.

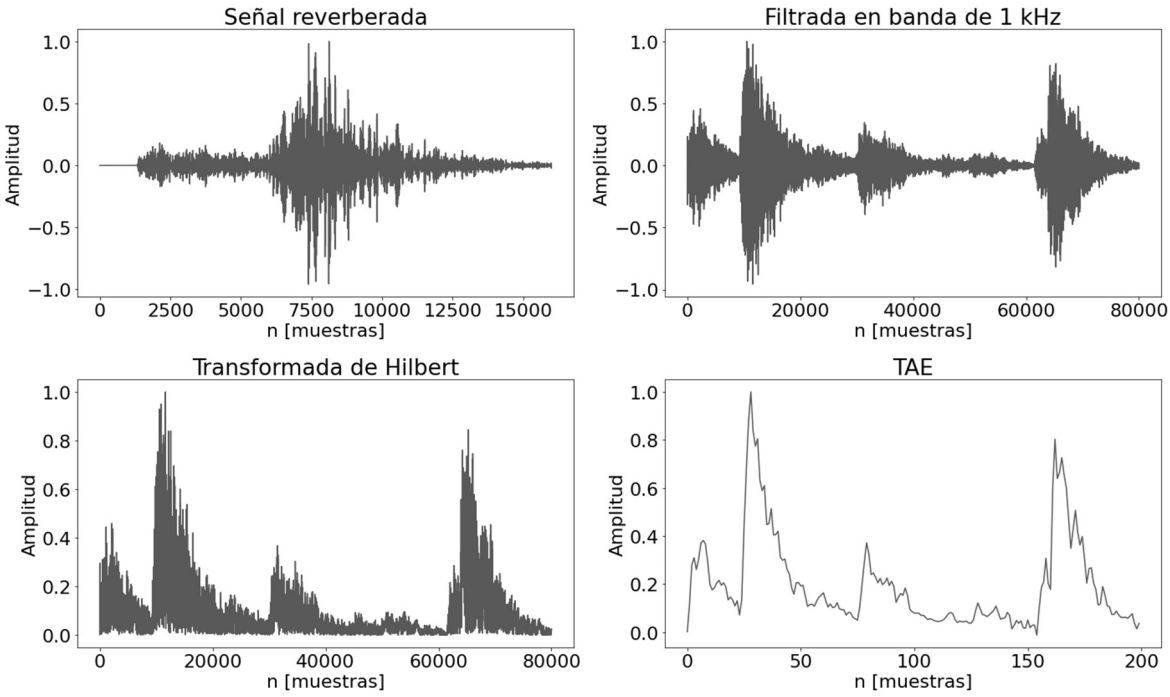

Figura 5. Pasos para la obtención de la TAE.

3.6 REDES NEURONALES ARTIFICIALES

3.6.1 EL PERCEPTRÓN MULTICAPA

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) y dieron pie a las investigaciones de lo que se conoce como aprendizaje profundo [25].

Este algoritmo simula el funcionamiento de aprendizaje de las neuronas biológicas a través de un modelo llamado neurona artificial o perceptrón.

Una red neuronal se organiza en capas, las cuales poseen una cierta cantidad de neuronas. Las salidas de las neuronas de una capa suelen constituir las entradas de las neuronas de la capa siguiente.

Una de las redes neuronales más estudiadas es el perceptrón multicapa [26]. 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.

Lo importante de este algoritmo es que, si la red tiene un número suficiente de neuronas, es capaz de representar cualquier función matemática.

3.6.2 ENTRENAMIENTO DE UNA RED NEURONAL

Los algoritmos de aprendizaje automático se entrenan a partir del procesamiento de datos. Cuando se trata de un modelo 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.

La encargada de encontrar esta configuración de la red 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 dicha función. Esta tarea es realizada por otra función denominada función de optimización. La misma 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. En base a este gradiente y al 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.

3.6.3 REDES NEURONALES CONVOLUCIONALES: CNN

Las redes neuronales convolucionales [27] se distinguen de otras redes neuronales por su rendimiento superior con entradas de señales de imagen, voz o audio. Tienen tres tipos principales de capas, que son:

- Capa convolucional
- Agrupación de capas (pooling layer)
- Capa totalmente conectada (FC)

La capa convolucional es el bloque central de una CNN, y es donde ocurre la mayoría de los cálculos. Requiere algunos componentes, que son datos de entrada, un filtro y un mapa de características. Si la entrada es una imagen en color, que se compone de una matriz de píxeles en 3D, la entrada tendrá tres dimensiones (alto, ancho y profundidad) que corresponden a RGB en una imagen. Además, esta capa cuenta con un detector de características, llamado kernel o filtro, que se mueve a través de los campos receptivos de la imagen, verificando si la característica está presente. Este proceso es el mismo al cálculo de la convolución, de ahí el nombre de estas redes.

El detector de características es una matriz bidimensional (2-D) de pesos, que representa parte de la imagen. El filtro se aplica a un área de la imagen y se calcula un producto punto entre los píxeles de entrada y el filtro. Este producto escalar luego se introduce en una matriz de salida. Luego, el filtro se mueve repitiendo el proceso hasta que el núcleo ha barrido toda la imagen. El resultado final de la serie de productos escalares de la entrada y el filtro se conoce como mapa de características, mapa de activación o característica convolucionada.

Para este tipo de modelo, hay tres hiperparámetros que afectan el tamaño del volumen de la salida que deben configurarse antes de que comience el entrenamiento de la red neuronal. Éstos incluyen:

- El **número de filtros**, el cual afecta la profundidad de la salida.
- El **stride**, que se refiere a la distancia que el núcleo se mueve sobre la matriz de entrada. Mientras más grande sea esta distancia, la salida será más pequeña.
- El **relleno de ceros**, el cual se suele utilizar cuando los filtros no se ajustan a la imagen de entrada.

Después de cada operación de convolución, la CNN aplica una transformación de unidad lineal rectificada (ReLU) al mapa de características, introduciendo la no linealidad en el modelo.

Por otro lado, la agrupación de capas, también conocida como reducción de resolución, lleva a cabo una disminución de la dimensionalidad, lo que minimiza la cantidad de parámetros en la entrada. Similar a la capa convolucional, la operación de agrupación barre un filtro a lo largo de toda la entrada, pero la diferencia es que este filtro no tiene ningún peso. En este caso, el kernel aplica una función de agregación a los valores dentro del campo receptivo, poblando la matriz de salida.

Existen dos tipos principales de agrupación:

- Agrupación máxima (max pooling): a medida que el filtro se mueve a través de la entrada, selecciona el píxel con el valor máximo para enviarlo a la matriz de salida.
- Agrupación promedio (average pooling): en este caso, cuando el filtro se mueve a través de la entrada, calcula el valor promedio dentro del campo receptivo para enviarlo a la matriz de salida.

Por último, la capa completamente conectada, como su nombre lo indica, es una capa donde cada nodo de la capa de salida se conecta directamente a un nodo de la capa anterior.

Esta realiza la tarea de clasificación en base a las características extraídas a través de las capas anteriores y sus diferentes filtros. Mientras que las capas convolucionales y de agrupación tienden a usar funciones ReLu, las capas FC generalmente aprovechan una función de activación softmax para clasificar las entradas de manera adecuada, produciendo una probabilidad de 0 a 1.

4. METODOLOGÍA

4.1 GENERACIÓN DE BASE DE DATOS

Para el entrenamiento de toda red neuronal es imprescindible contar con una base de datos extensa. La misma debe poder representar, de la mejor manera posible, el fenómeno que se quiere modelar. Además, se debe poder asegurar que todas las instancias que componen la base de datos compartan las mismas características: formato de audio, duración, profundidad de bits y frecuencia de muestreo.

Para esta investigación se requieren audio de voz reverberados, y los valores de los parámetros acústicos del recinto que proviene dicha reverberación. Para obtener esto, se utilizaron audios anecoicos y se reverberaron utilizando diversas respuestas al impulso, tal como se expresa en la ecuación 1. Para esto, se debe asumir que los recintos son sistemas LTI y por lo tanto, las señales reverberadas pueden obtenerse convolucionando las señales anecoicas con la repuesta al impulso del recinto.

La base de datos generada se divide en los siguientes conjuntos de datos: entrenamiento, validación y prueba. Los conjuntos de validación y prueba se utilizan para verificar que el sistema sea capaz de generalizar a condiciones no vistas durante el entrenamiento. Por esto, se utilizan respuestas al impulso distintas en cada conjunto generado.

En concreto, se generaron un total de 29000 audios de 5 s de duración cada uno y una frecuencia de muestreo de 16000 Hz, lo que equivaldría a una base de datos de poco más de 40 horas de duración. De esta, se separó de forma aleatoria un 72% para la etapa de entrenamiento, un 20\% para la etapa de verificación y un 8% para validación.

Los audios se generaron a través de reverberar 10 señales de voz anecoicas, 5 hombres y 5 mujeres, de la base de datos de la ACE [12] con RIRs sintéticas con tiempos de reverberación que iban desde los 0,2 a 3 s en pasos de 0,1 s. Dichas respuestas al impulso se generaron mediante lo expresado en la ecuación 11.

Finalmente, cada audio generado se empareja en la etapa de entrenamiento de la red junto a su valor objetivo. En este caso, para cada audio de entrada, se espera como valor objetivo (y) el valor del descriptor T30 de la RIR que se utilizó para reverberarlo.

4.2 MODELO PROPUESTO

Para la obtención ciega de parámetros acústicos, se propone un método inspirado en la investigación de Kendrick et al. [9] sobre la obtención del tiempo de reverberación de una sala a partir de la curva de decaimiento de un audio reverberado y el modelo propuesto en la investigación de Duangpummet et. al. [8].

En concreto, para cada audio reverberado, se filtra el mismo por las bandas de octava de 125, 250, 500, 1000, 2000, 4000 y 8000 Hz, se calcula su TAE por cada banda, tal como se vio en la sección 3.5, y se ingresa el mismo a una red neuronal convolucional cuyo valor de salida esperado es el T30 de la RIR con la que se reverberó el audio.

A modo ilustrativo, en la Figura 6 se puede observar el diagrama en bloque del modelo propuesto para esta investigación.

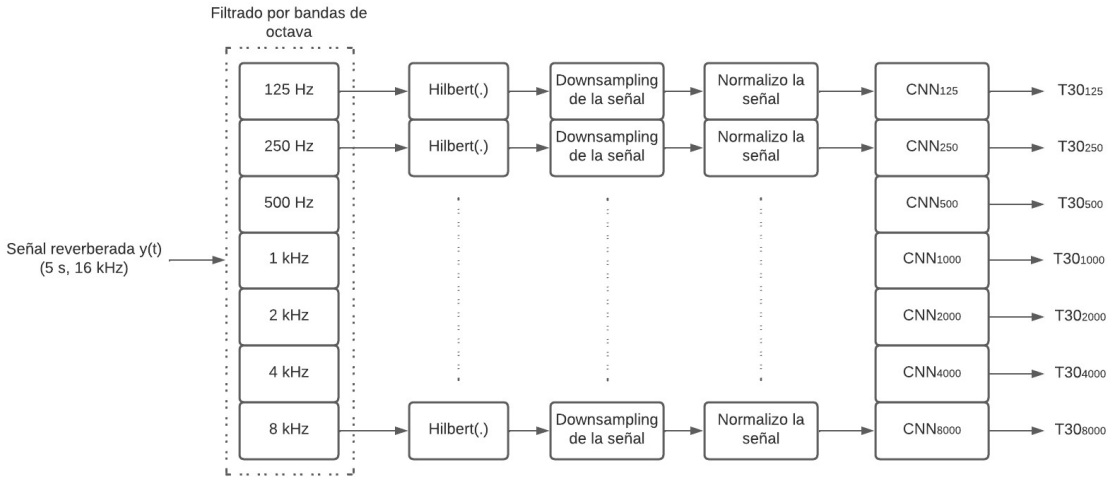

Figura 6. Diagrama en bloques del modelo propuesto.

En cuanto a la red neuronal, se crearon 7 CNN de una única dimensión para tratar de estimar los valores de T30 a partir de las características de las TAEs de entrada. Estas redes son idénticas y se utiliza una por cada banda.

Cada modelo cuenta con 4 capas convolucionales. La capa 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 capa de salida, la cual trata de estimar el valor del T30 a través de conocer su valor esperado. La compilación del modelo se hace a través del optimizador de Adam.

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

Tabla 2. Arquitectura de red propuesta.

N°	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}	16 filtros, kernel=5, ReLu
6	Pooling	max pooling, pool_size=2
7	Normalización	batch normalization
8	Dropout	40%
9	Conv1D^{3era}	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	1 salida (T30), ReLu
14	Regresión de salida	Error cuadrático medio (MSE)

4.3 EVALUACIÓN DEL MODELO

Para evaluar el desempeño de la red, una vez entrenada, se utilizan los pesos sinápticos obtenidos para cada capa del modelo y se realiza una predicción del tiempo de reverberación a partir de una TAE desconocida y se la compara con su valor real.

En concreto, se realiza esta comparación con todo el conjunto de datos de prueba y se calcula el coeficiente de correlación de Spearman [28] para determinar qué tan eficiente es la red para estimar los valores. Esto se repite para las 7 bandas de estudio.

5. RESULTADOS Y ANÁLISIS

En esta sección se presentan los resultados obtenidos durante el entrenamiento de las 7 redes neuronales. Cada una de ellas cuenta con la misma arquitectura, pero están destinadas al cálculo de los descriptores en distintas bandas de frecuencia. En concreto, se entrenaron para las bandas de octava de 125, 250, 500, 1000, 2000, 4000 y 8000 Hz respectivamente.

Se utilizó una base de datos de 29000 audios de 5 segundos de duración cada uno y con 16 kHz de frecuencia de muestreo. Para cada caso, los audios se filtraban en bandas de octava según la red que se estaba por entrenar.

5.1 ENTRENAMIENTO DE LOS MODELOS

Los modelos se entrenaron en un total de 100 épocas, usando el error cuadrático medio como función de costo o loss. Del total de la base de datos, se destinó un 80% para el entrenamiento, del cual, además, se reservó un 10% para la etapa de validación.

Con esta elección de épocas, el algoritmo tiene las suficientes iteraciones para poder reducir su función de costo pero sin llegar a memorizar los datos. Esto se puede observar en la Figura 7, 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).

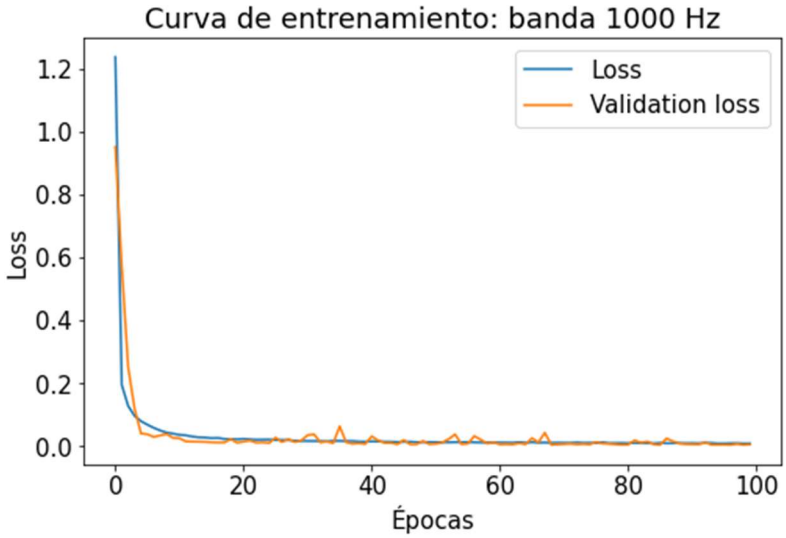

Figura 7. Curvas de entrenamiento de la red de la banda de 1 kHz.

Se decidió mostrar en esta etapa el entrenamiento para la banda de 1 kHz, los otros 6 casos restantes se pueden observar en el Anexo.

Con esto, al menos en la parte de entrenamiento, se observó que el modelo propuesto es capaz de estimar el tiempo de reverberación a partir de la TAE de un audio de voz reverberado. En la siguiente sección se evalúa el desempeño del mismo pero ahora con audios que la red no usó en su entrenamiento.

5.2 EVALUACIÓN DEL MODELO CON AUDIOS DESCONOCIDOS

En esta ocasión, se busca determinar la eficiencia del modelo para predecir los valores de tiempo de reverberación de una sala probándolos con el 20% restante de la base de datos que no se utilizó durante el entrenamiento.

En concreto, se calculó el coeficiente de correlación de Spearman (y su p_valor asociado) tomando los valores que pudo predecir la red en contraste con los valores de T30 reales.

Los resultados por banda de frecuencia se pueden observar en la tabla 3 y, a su vez, a modo ilustrativo se pueden observar las rectas de regresión en la Figura 8.

Tabla 3. Coeficiente de correlación de Spearman entre las predicciones y el valor real de T30 por cada banda.

Banda [Hz]	r	p_valor
125	0.984	<0.01
250	0.988	<0.01
500	0.993	<0.01
1000	0.996	<0.01
2000	0.997	<0.01
4000	0.996	<0.01
8000	0.997	<0.01

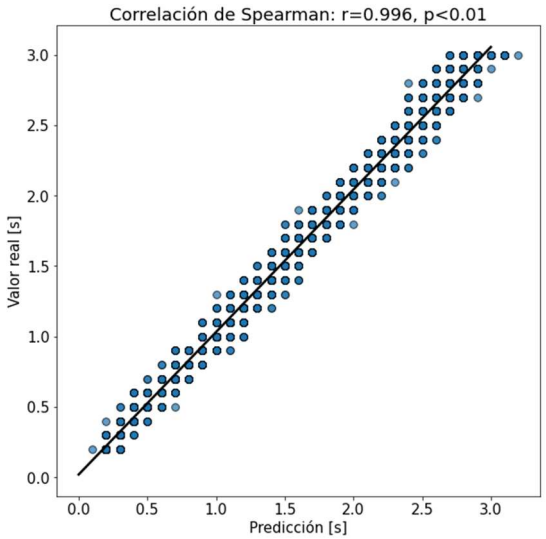

Figura 8. Rectas de regresión entre los valores predichos y los reales para la banda de 1 kHz.

Las rectas de regresión de las demás bandas se pueden encontrar en el Anexo.

En todos los casos, la red tuvo una precisión mayor al 98% para predecir el valor del tiempo de reverberación.

Se observó que la precisión de la estimación aumenta en frecuencias altas. Inicialmente, se podría intuir que la precisión debería ser mejor en las frecuencias donde predomina la voz (ya que las estimaciones se hacen a partir de audios grabados), pero esto no es necesariamente así. En este entrenamiento en particular, la cantidad de contenido

frecuencial no es de importancia, sino la cantidad de decaimiento del nivel en la envolvente de los TAE.

6. CONCLUSIONES

Aquí se aplicarán los métodos estadísticos que nos permitan concluir los resultados y proyectar diferentes facetas de los datos obtenidos. Como las pruebas no se realizarán todavía, deben plantearse los métodos elegidos y explicarlos.

7. LÍNEAS FUTURAS DE INVESTIGACIÓN

En las conclusiones del Plan de Investigación, debe plantearse cómo será la exposición de los resultados y qué es lo que se espera obtener en resumen de las pruebas que se realicen.

BIBLIOGRAFÍA

[1] V. G. Escobar y J. B. Morillas, Analysis of intelligibility and reverberation time recommendations in educational rooms, Applied Acoustics 96, 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, 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, 41-44 (2010).
[4] ISO 3382-1:2009, Acoustics. Measurement of room acoustic parameters. Part 1: Performance spaces, International Organization for Standardization, Geneva (2009).
[5] IEC 60268-16:2020, Sound system equipment. Part 16: Objective rating of speech intelligibility by speech transmission index, International Electrotechnical Commission, Geneva (2021).
[6] A. V. Oppenheim y A. S. Ian, Signals and systems, edited by Prentice-Hall, 2nd edition, (1996) Págs. 74-137.
[7] N. M. Papadakis y G. E. Stavroulakis, Review of Acoustic Sources Alternatives to a Dodecahedron Speaker, MDPI, 1-32 (2019).
[8] S. Duangpummet, J. Karnjana, W. Kongprawechnon y M. Unoki, Blind Estimation of Room Acoustic Parameters and Speech Transmission Index using MTF-based CNNs, ELSEVIER, 1-12 (2021).
[9] 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, 1-11 (2007).
[10] P. Kendrick, F. F. Li y T. J. Cox, Monaural room acoustic parameters from music and speech, The Journal of the Acoustical Society of America, 1-11 (2008).
[11] THE ACE CHALLENGE, Extraído el 18 de abril del 2022, http://www.ee.ic.ac.uk/naylor/ACEweb/index.html
[12] 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, 1-10 (2016).
[13] 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, 1-5 (2015).
[14] 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, 1-5 (2015).

[15] 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, 1-5 (2015).
[16] N. J. Bryan, Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation, Adobe Research, 1-5 (2019).
[17] S. W. Smith, The scientist and engineer's guide to digital signal processing, edited by Calif: California Technical Pub, 6th edition, (1997) Págs. 277-284.
[18] Y.-W. Liu, Fourier Transform Applications, edited by InTech, 1st edition, (2012) Págs. 291-300.
[19] M. R. Schroeder, New method of measuring reverberation time, The Journal of the Acoustical Society of America, vol. 37, n.° 6, 1187-1188 (1965).
[20] A. Lundeby, T. E. Vigran, H. Bietz y M. Vorländer, Uncertainties of measurements in room acoustics, Acta Acustica united with Acustica, vol. 81, n.° 4, 344-355 (1995).
[21] A. Lundeby, T. E. Vigran, H. Bietz y M. Vorländer, Uncertainties of measurements in room acoustics, Acta Acustica united with Acustica, vol. 81, n.° 4, 344-355 (1995).
[22] H. J. Steeneken y T. Houtgast, A physical method for measuring speech-transmission quality, The Journal of the Acoustical Society of America, vol. 67, n.° 1, 318-326 (1980).
[23] H. Kuttruff, Room acoustics, edited by Crc Press, 6th edition, (2016).
[24] M. R. Schroeder, Modulation transfer functions: Definition and measurement, Acta Acustica united with Acustica, vol. 49, n.° 3, 179-182 (1981).
[25] I. A. Basheer y M. Hajmeer, Artificial neural networks: fundamentals, computing, design, and application, Journal of microbiological methods, vol. 43, n.° 1, 3-31 (2000).
[26] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural networks, vol. 4, n.° 2, 251-257 (1991).
[27] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems, edited by O'Reilly Media, Inc., 2nd edition, Págs. 431-461 (2019).
[28] L. Myers y M. J. Sirois, Spearman correlation coefficients, differences between, Encyclopedia of statistical sciences, vol. 12, 2004.

ANEXO I.

A 1. Curvas de entrenamiento de los modelos:

En este apartado se pueden observar las curvas de entrenamiento de las 7 redes neuronales entrenadas. A saber, en cada caso, la arquitectura es la misma pero los datos de entrenamiento están filtrados con filtros pasabanda centrados en las frecuencias de 125, 250, 500, 1000, 2000, 4000 y 8000 Hz.

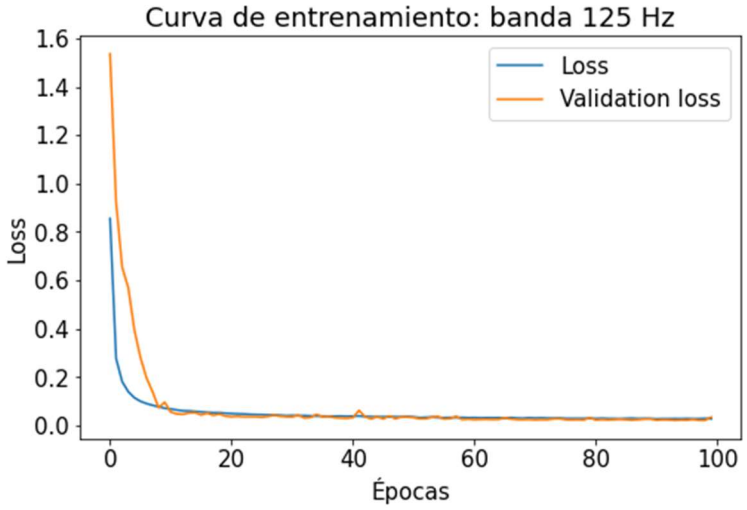

Figura 1. Curvas de entrenamiento de la red de la banda de 125 Hz

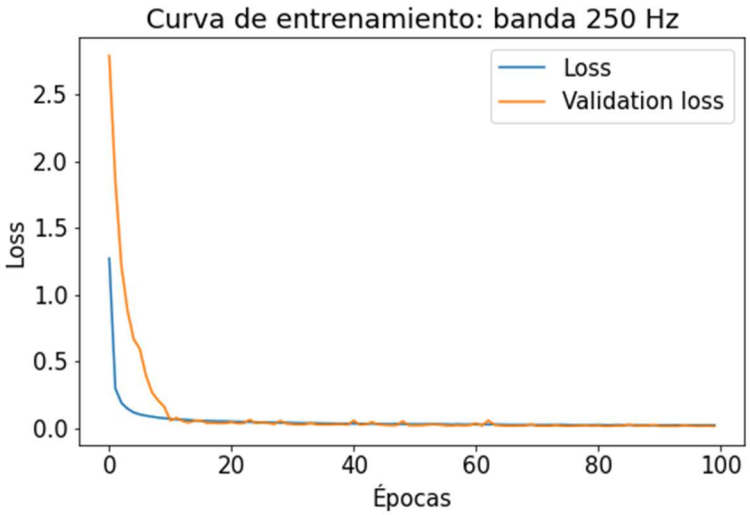

Figura 2. Curvas de entrenamiento de la red de la banda de 250 Hz

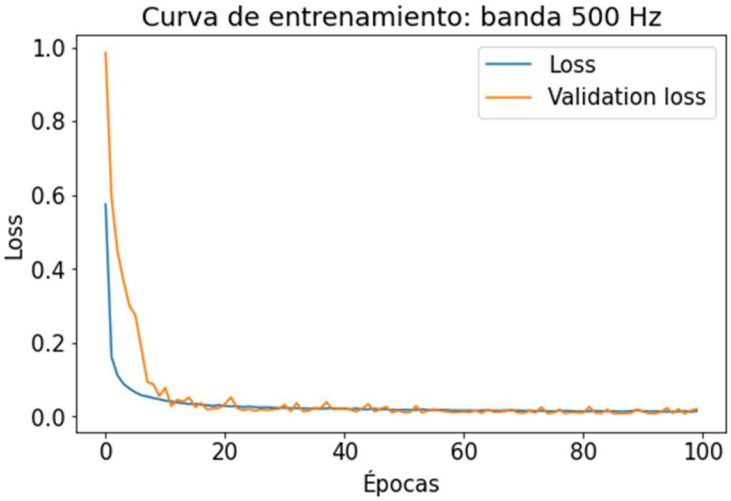

Figura 3. Curvas de entrenamiento de la red de la banda de 500 Hz

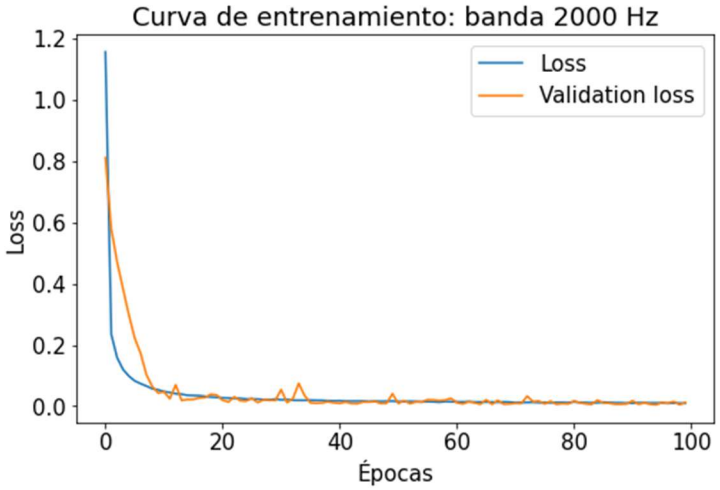

Figura 4. Curvas de entrenamiento de la red de la banda de 2 kHz

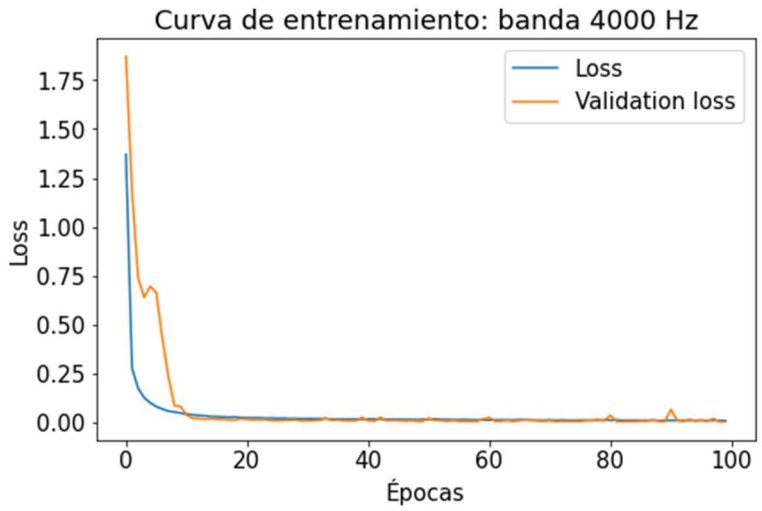

Figura 5. Curvas de entrenamiento de la red de la banda de 4 kHz

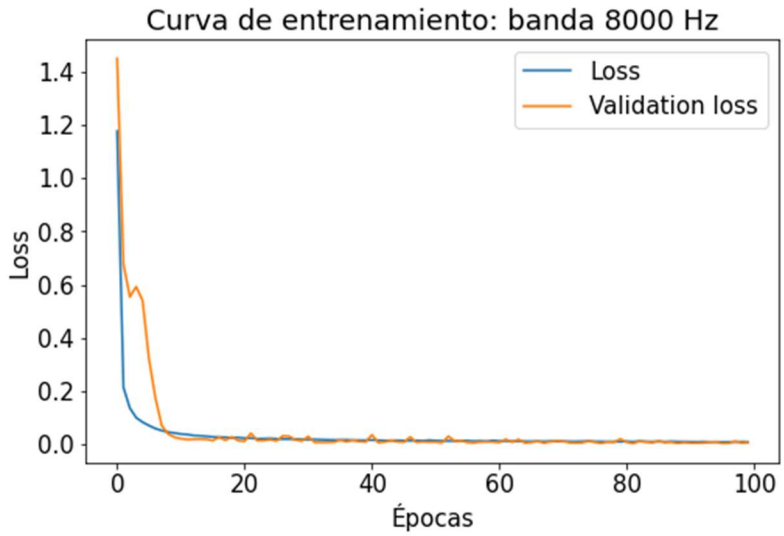

Figura 6. Curvas de entrenamiento de la red de la banda de 8 kHz

A 2. Rectas de regresión de los resultados obtenidos:

En este apartado se pueden observar las rectas de regresión obtenidos en el testeo de las 7 redes neuronales. A saber, para cada una, los datos de prueba están filtrados con filtros pasabanda centrados en las frecuencias de 125, 250, 500, 1000, 2000, 4000 y 8000 Hz.

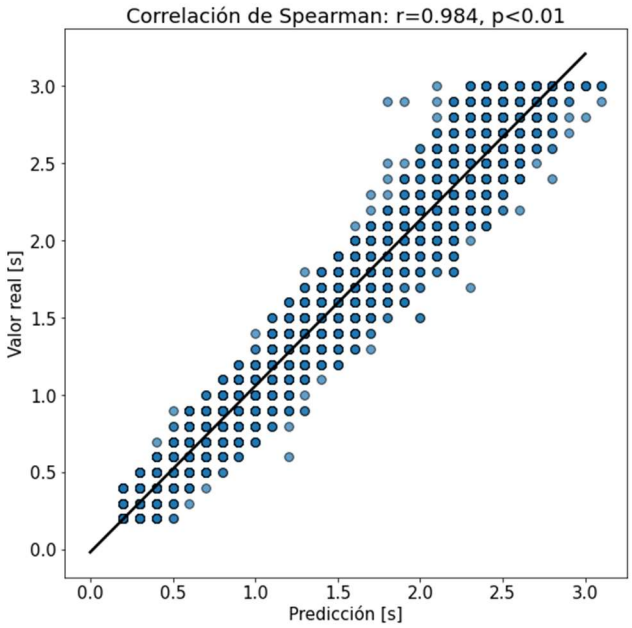

Figura 7. Rectas de regresión entre los valores predichos y los reales para la banda de 125 Hz.

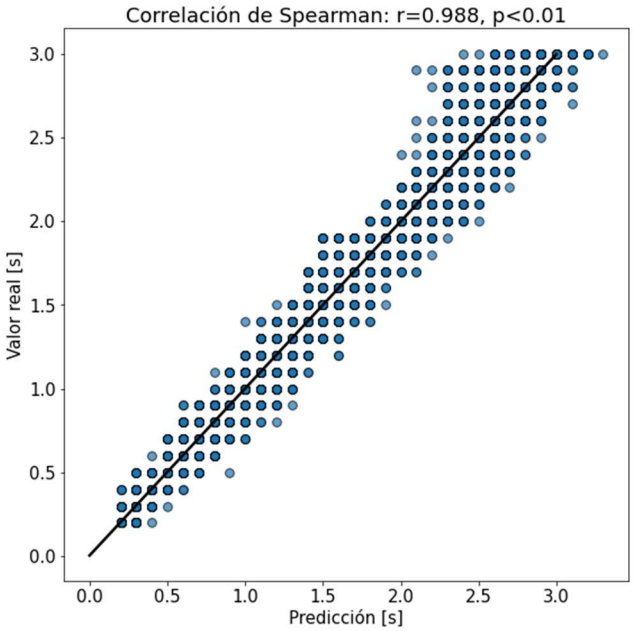

Figura 8. Rectas de regresión entre los valores predichos y los reales para la banda de 250 Hz.

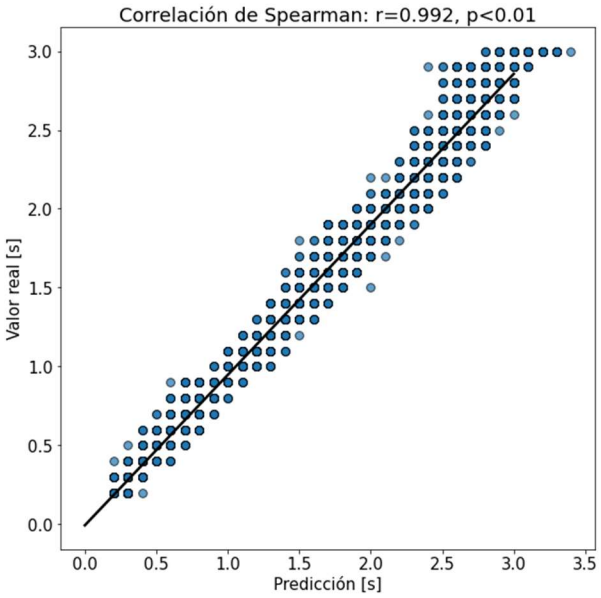

Figura 9. Rectas de regresión entre los valores predichos y los reales para la banda de 500 Hz.

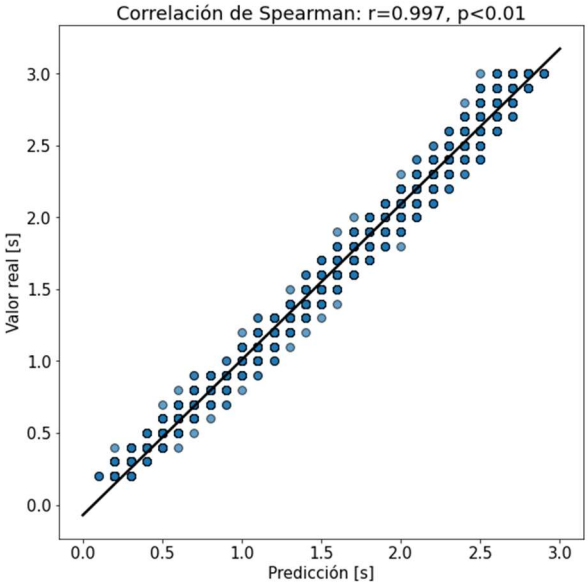

Figura 10. Rectas de regresión entre los valores predichos y los reales para la banda de 2 kHz.

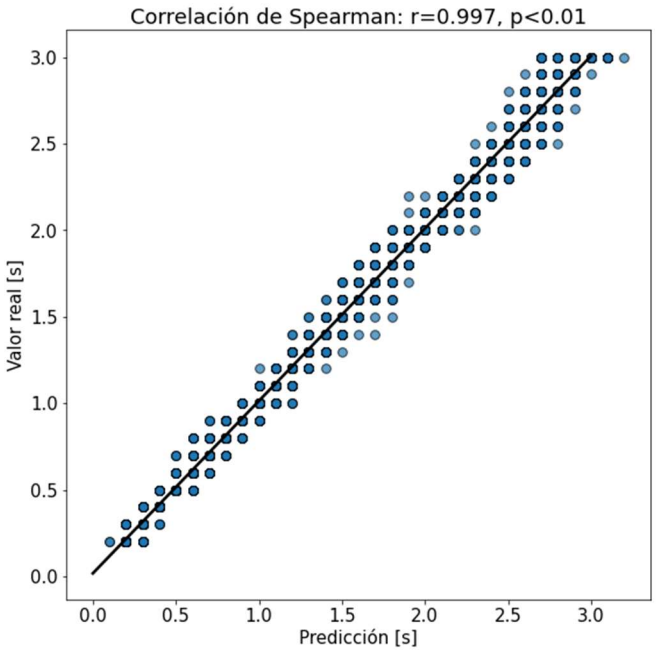

Figura 11. Rectas de regresión entre los valores predichos y los reales para la banda de 4 kHz.

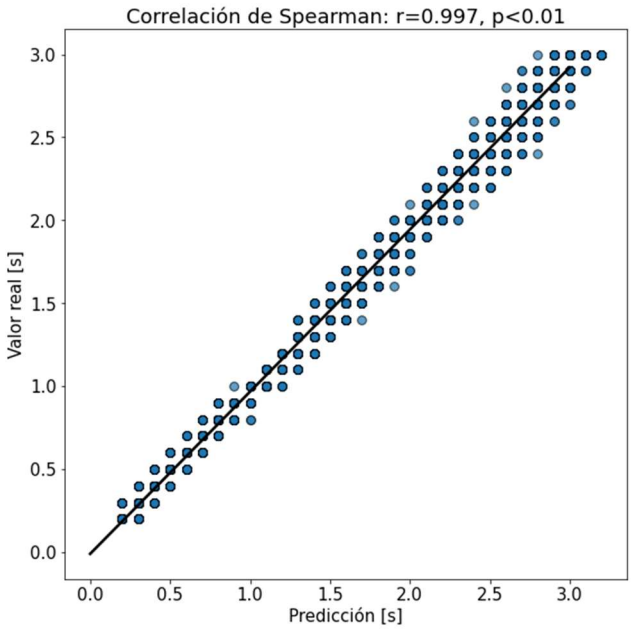

Figura 12. Rectas de regresión entre los valores predichos y los reales para la banda de 8 kHz.