Escola Tècnica Superior d'Enginyeria ETSE-UV

Programa de Doctorat en Tecnologies de la Informació, Comunicacions i Computació.

Vniversitat de València

Tesis Doctoral

Monitorización IoT para la caracterización acústica del paisaje sonoro mediante aprendizaje profundo.

Autor:

Jesús López Ballester

Directores:

Jaume Segura García

Santiago Felici Castell

Septiembre, 2022

Escola Tècnica Superior d'Enginyeria ETSE-UV

Programa de Doctorat en Tecnologies de la Informació, Comunicacions i Computació.

Tesis Doctoral

Monitorización IoT para la caracterización acústica del paisaje sonoro mediante aprendizaje profundo.

Autor: Jesús López Ballester

Directores: Jaume Segura García y Santiago Felici Castell

Septiembre, 2022

La presente Tesis Doctoral es parte del proyecto de I+D+i BIA2016-76957-C3-1-R y se ha realizado mediante la ayuda BES-2017-082340, financiados por el Ministerio de Ciencia e Innovación, la Agencia Estatal de Investigación, MCIN/AEI/10.13039/501100011033/ y por los Fondos de la Unión Europea "FEDER Una manera de hacer Europa" y "FSE Invierte en tu futuro".

A mi familia, la de fuera y la de dentro.

Agradecimientos

El desarrollo de esta Tesis Doctoral se ha realizado durante varios años, y al incluir un gran contenido práctico de diferentes disciplinas, ha propiciado que muchas personas hayan aportado un conocimiento, tiempo y esfuerzo sin el que sin duda no se habría podido realizar, y a las que me gustaría agradecer su aportación.

Me gustaría agradecer a mis directores de Tesis Jaume y Santi su apoyo y orientación constantes, su pasión por sacar el máximo provecho a todas las facetas de los resultados extraídos y la aportación de nuevas ideas, posibles soluciones y líneas de investigación de manera continua. Sin sus esfuerzos de revisión minuciosos, sus aportaciones y críticas constructivas, ni la investigación realizada ni las contribuciones publicadas tendrían el nivel de calidad que tienen. Del mismo modo me gustaría agradecer a Máximo su ayuda en todos los ámbitos de esta Tesis, pues su guía y apoyo se han manifestado en los resultados obtenidos durante el desarrollo de la misma. Su capacidad de análisis y crítica han contribuido a realizar todos los procesos de forma rigurosa, precisa y documentada, logrando resultados de gran valor.

Junto a los ya mencionados, en lo que respecta a docencia, me gustaría agradecer especialmente la ayuda de Esther, que ha sido inestimable a la hora de desarrollar, preparar e impartir clases. Sin sus conocimientos y la gran cantidad de tiempo ayudándome a "pelear" con materiales, plataformas y numerosas tareas de gestión mi labor docente hubiera sido infinitamente más compleja. Gracias a Sandra y Máximo también por esos ratos en el café arreglando el mundo, pues me han ayudado mucho a abordar mejor problemas y soluciones, creciendo como persona. En el punto de tormenta de ideas rodeados de café o comida, tendría que mencionar a más de medio departamento de informática con los que he compartido muchos ratos e ideas aprendiendo de todo un poco, gracias a todos los que han participado de esos momentos.

A todos los amigos que me aportaban su punto de vista escuchando batallitas, permitiéndome ver las cosas desde fuera con otra actitud y sobre todo haciéndome desconectar todos los momentos posibles. A mis padres, por estar ahí pese a cualquier dificultad. A mi mujer Noelia por hacer junto a mí el "viaje"de esta Tesis y darme todo el apoyo que ha podido en cada momento. A mi hija Valeria, por aparecer en medio de toda esta locura haciendo que lo vea todo con una luz distinta.

Resumen

El sonido que nos rodea forma lo que denominamos paisaje sonoro o *soundscape* y ha ido ganando interés con los años al estudiarse su efecto bajo el prisma de diferentes campos de investigación tanto biológicos como físicos, como la medicina, la antropología, la arquitectura o la ecología. El paisaje sonoro es fundamental para el confort, la vida y al fin y al cabo la salud de los habitantes de un entorno, por lo que analizarlo y describirlo de una manera objetiva mediante parámetros cuantificables tiene una gran importancia. Para ello se definen diversos parámetros acústicos que se emplean en función del soundscape que se desee estudiar o del tipo de análisis a realizar.

Algunos de estos parámetros suelen ser conocidos como puede ser la Sonoridad (*Loudness*), que emplearíamos para analizar el nivel de presión sonora del sonido presente en una zona de una ciudad o la Reverberación (*Reverberation Time*), que emplearíamos para estudiar cuánto tiempo percibimos el sonido en una sala debido a las reflexiones, otros en cambio nos son menos familiares, como el caso de del índice de inteligibilidad del habla (*Speech Intelligibility Index* o *SII*), en el caso de querer analizar cómo se perciben las palabras en diferentes zonas de una sala. Dependiendo de las características que se deseen analizar del sonido, existen diversos parámetros como pueden ser además de los mencionados, la Intensidad de Fluctuación o *Fluctuation Stregth*, la Claridad musical o *Clarity* 80 o la Rugosidad (*Roughness*) entre otros. Estos parámetros acústicos, individualmente aportan una información concreta del sonido analizado, pero en conjunto proporcionan una descripción muy amplia del soundscape analizado, por lo que es sumamente interesante la implementación de un sistema capaz de evaluar diversos parámetros acústicos de forma simultánea y obtener en un mismo análisis una descripción lo más completa posible. Al igual que tener un único valor de temperatura para una ciudad al completo, no nos permite saber dónde se ubican las zonas más cálidas y las más frías, a nivel acústico, es esencial monitorizar diferentes zonas para poder realizar un análisis correcto y saber cómo varían los parámetros acústicos en las mismas. Para este cometido nacen las redes inalámbricas de sensores acústicos o *Wireless Acoustic Sensors Networks* (*WASNs*), que nos permiten monitorizar diferentes puntos de un soundscape simultáneamente y realizar análisis extensos y complejos de forma sencilla y rápida.

Las WASNs están ganando protagonismo cada día en la monitorización de entornos, más todavía si cabe con la mejora de la conectividad de los dispositivos y la irrupción del concepto del Internet de las Cosas o Internet of Things (*IoT*) en el mundo de los dispositivos móviles y las redes de sensores. En la mayoría de casos las WASNs están formadas por dispositivos de bajo coste y por lo tanto recursos limitados, por lo que la velocidad y eficiencia de cálculo es esencial. Los sistemas IoT pueden estar formados por dispositivos de naturaleza más variada además de los nodos típicos de una WASN, pero en ellos es igual de importante la velocidad de cálculo, permitiendo dedicar tiempo a realizar otras tareas como el almacenamiento de datos en la nube o la implementación de sistemas de representación de los mismos. En muchos casos

los nodos que forman la WASN o el sistema IoT están alimentados por baterías, por lo que una vez más la eficiencia y velocidad de cálculo son esenciales. La Inteligencia Artificial (IA) juega un papel fundamental en este punto ya que permite automatizar y simplificar tareas y cálculos complejos. Empleando redes neuronales convolucionales (*Convolutional Neural Networks*, $CNNs$) se puede obtener uno o varios valores a su salida en función de los datos presentes a la entrada de una forma mucho más rápida que mediante el cálculo directo. Embebiendo este tipo de redes neuronales en un sistema IoT, se da lugar al concepto AI-IoT, habilitando dispositivos de recursos limitados para realizar cálculos complejos de una manera sencilla, rápida y con un consumo reducido de energía.

En esta Tesis se describe el uso de dos conjuntos de parámetros acústicos, uno para evaluar la molestia psico-acústica en entornos generalmente amplios y otro para evaluar el comportamiento acústico general de una sala y la inteligibilidad del habla en la ella. Debido a la complejidad de los cálculos necesarios se ha hecho uso de redes neuronales convolucionales para acelerar los mismos y poder realizarlos dentro de un nodo de una WASN. Las CNNs se han diseñado de manera que minimicen el error en predicción al máximo manteniendo una elevada velocidad de cálculo. Se han puesto a prueba con diferentes datasets incluyendo señales reales adquiridas con diferentes dispositivos, para verificar el correcto funcionamiento en entornos reales bajo la influencia de cualquier agente externo. Gracias a esto se ha diseñado y desplegado un sistema IoT que hace uso de las redes neuronales diseñadas para la monitorización de los parámetros acústicos de forma precisa, sencilla y rápida. El sistema AI-IoT es capaz de almacenar los datos de forma local o en la nube y de predecir los parámetros a partir de señales de audio sin procesar o *RAW* mucho más rápido que empleando cálculo directo, por lo que sumado al uso de baterías en los nodos y a la comunicación inalámbrica, permite analizar un soundscape de una forma mucho más sencilla y rápida que ningún sistema empleado hasta ahora.

Palabras clave

Internet de las cosas, parámetros acústicos de sala, parámetros psico-acústicos, redes de sensores acústicos inalámbricos, redes neuronales convolucionales, regresión, soundscape.

Resum

Els sons que ens envolten formen el que denominem paisatge sonor o *soundscape* i ha anat guanyant interés amb els anys en estudiar-se el seu efecte sota el prisma de diferents camps d'investigació tant biològics com físics, com la medicina, l'antropologia, l'arquitectura o l'ecologia. El paisatge sonor és fonamental per al confort, la vida i al cap i a la fi la salut dels habitants d'un entorn, per la qual cosa analitzar-lo i descriure'l d'una manera objectiva mitjançant paràmetres quantificables té una gran importància. Per a això es defineixen diversos paràmetres acústics que s'empren en funció del soundscape que es desitge estudiar o del tipus d'anàlisi a realitzar.

Alguns d'aquests paràmetres solen ser coneguts com pot ser la Sonoritat (*Loudness*), que empraríem per a analitzar el nivell de pressió sonora del so present en una zona d'una ciutat o la Reverberació (*Reverberation Time*), que empraríem per a estudiar quant temps percebem el so en una sala a causa de les reflexions, uns altres en canvi ens són menys familiars, com el cas de de l'índex d'intel·ligibilitat de la parla (*Speech Intelligibility Index* o *SII*), en el cas de voler analitzar com es perceben les paraules en diferents zones d'una sala. Depenent de les característiques que es desitgen analitzar del so, existeixen diversos paràmetres com poden ser a més dels esmentats, la Intensitat de Fluctuació o *Fluctuation Stregth*, la Claredat Musical o *Clarity* 80 o la Rugositat (*Roughness*) entre altres. Aquests paràmetres acústics, individualment aporten una informació concreta del so analitzat, però en conjunt proporcionen una descripció molt àmplia del soundscape analitzat, per la qual cosa és summament interessant la implementació d'un sistema capaç d'avaluar diversos paràmetres acústics de manera simultània i obtindre en una mateixa anàlisi una descripció el més completa possible. Igual que tindre un únic valor de temperatura per a una ciutat al complet, no ens permet saber on se situen les zones més càlides i les més fredes, a nivell acústic, és essencial monitorar diferents zones per a poder realitzar una anàlisi correcta i saber com varien els paràmetres acústics en aquestes. Per a aquesta comesa naixen les xarxes sense fils de sensors acústics o *Wireless Acoustic Sensors Networks* (*WASNs*), que ens permeten monitorar diferents punts d'un soundscape simultàniament i realitzar anàlisis extenses i complexos de manera senzilla i ràpida.

Les WASNs estan guanyant protagonisme cada dia en el monitoratge d'entorns, més encara si cap amb l'increment de la connectivitat dels dispositius i la irrupció del concepte de la Internet de les Coses o Internet of Things (*IoT*) en el món dels dispositius mòbils i les xarxes de sensors. En la majoria de casos les WASNs estan formades per dispositius de baix cost i per tant recursos limitats, per la qual cosa la velocitat i eficiència de càlcul és essencial. Els sistemes IoT poden estar formats per dispositius de naturalesa més variada a més dels nodes típics d'una WASN, però en ells és igual d'important la velocitat de càlcul, permetent dedicar temps a fer altres tasques com l'emmagatzematge de dades en el núvol o la implementació de sistemes de representació d'aquests. En molts casos els nodes que formen la WASN o el sistema IoT estan alimentats per bateries, per la qual cosa una vegada més l'eficiència i velocitat

de càlcul són essencials. La Intel·ligència Artificial (IA) juga un paper fonamental en aquest punt ja que permet automatitzar i simplificar tasques i càlculs complexos. Emprant xarxes neuronals convolucionals (*Convolutional Neural Networks*, $CNNs$) es pot obtindre un o diversos valors a la seua eixida en funció de les dades presents a l'entrada d'una forma molt més ràpida que mitjançant el càlcul directe. Embevent aquest tipus de xarxes neuronals en un sistema IoT, es dona lloc al concepte AI-IoT, habilitant dispositius de recursos limitats per a realitzar càlculs complexos d'una manera senzilla, ràpida i amb un consum reduït d'energia.

En aquesta Tesi es descriu l'ús de dos conjunts de paràmetres acústics, un per a avaluar la molèstia psico-acústica en entorns generalment amplis i un altre per a avaluar el comportament acústic general d'una sala i la intel·ligibilitat de la parla en ella. A causa de la complexitat dels càlculs necessaris s'ha fet ús de les xarxes neuronals convolucionals per a accelerar els mateixos i poder realitzar-los dins d'un node d'una WASN. Les CNNs s'han dissenyat de manera que minimitzen l'error en predicció al màxim mantenint una elevada velocitat de càlcul. S'han posat a prova amb diferents datasets incloent senyals reals adquirits amb diferents dispositius, per a verificar el correcte funcionament en entorns reals sota la influència de qualsevol agent extern. Gràcies a això s'ha dissenyat i desplegat un sistema IoT que fa ús de les xarxes neuronals dissenyades per al monitoratge dels paràmetres acústics de manera precisa, senzilla i ràpida. El sistema AI-IoT és capaç d'emmagatzemar les dades de manera local o en el núvol i de predir els paràmetres a partir de senyals d'àudio sense processar o RAW molt més ràpid que emprant càlcul directe, per la qual cosa sumat a l'ús de bateries en els nodes i a la comunicació sense fil, permet analitzar un soundscape d'una forma molt més senzilla i ràpida que cap sistema emprat fins ara.

Paraules clau
Internet de les coses, paràmetres acústics de sala, paràmetres psico-acústics, regressió, soundscape, xarxes de sensors acústics sense fils, xarxes neuronals convolucionals.

Abstract

The sound that surrounds us forms what we call *soundscape* and has been gaining interest in research over the years as its effect has been studied under the prism of different fields of research, both biological and physical, such as medicine, anthropology, architecture or ecology. The soundscape is fundamental to the well-being, quality of life and, in summary, the health of the inhabitants of an environment. Therefore, it is very important to analyze and describe it objectively by means of quantifiable parameters. For this purpose, diverse acoustic parameters are defined and used depending on the soundscape to be studied or the type of analysis to be performed.

Some of these parameters are usually known, such as *Loudness*, which we would use to analyze the sound pressure level of sound present in an area of a city, or *Reverberation Time*, which we would use to study how long we perceive the sound in a room due to reflections. Others are less familiar to us, such as the *Speech Intelligibility Index* or SII, in the case of wanting to analyze how words are perceived in different areas of a room. Depending on the characteristics of the sound to be analyzed, there are several parameters such as *Fluctuation Stregth*, *Clarity* 80 or *Roughness* among others. These acoustic parameters, individually provide specific information about the analyzed sound, but together they provide a wide description of the analyzed soundscape, so it is very interesting to implement a system with the capability to evaluate several acoustic parameters simultaneously and to obtain in the same analysis a description as complete as possible. These acoustic parameters individually provide specific information about the analyzed sound, but together they provide a very broad description of the analyzed soundscape, so it is extremely interesting to implement a system capable of evaluating several acoustic parameters simultaneously and obtain a description as complete as possible in a single analysis.

Just as having a single temperature value for an entire city does not allow us to know where the warmest and coldest areas are located, in acoustic analysis it is essential to monitoring different areas in order to perform a precise analysis and to know how the acoustic parameters vary in these areas. For this purpose, *Wireless Acoustic Sensors Networks* ($WASNs$) are created, which allow us to carry out measurements at different points of a soundscape simultaneously and to perform extensive and complex analyses in a simple and fast way.

WASNs are becoming more popular every day in environment monitoring, even more so with the improvement of device connectivity and the emergence of the Internet of Things (IoT) concept in the world of mobile devices and sensor networks. In most cases WASNs are composed of low-cost devices and therefore limited resources, so speed and computational efficiency are essential. IoT systems may consist of devices of a more varied nature compared to the typical nodes of a WASN, but computational speed is also very important in them, as they may spend time performing other tasks, such as storing data in the cloud or implementing data representation systems. In many cases the nodes that form the WASN or IoT system are also powered by batteries, so once again efficiency and computational speed are

essential.

Artificial Intelligence (AI) plays a fundamental role here, since it allows to automate and simplify complex tasks and calculations. By employing *Convolutional Neural Networks* ($CNNs$) one or several values can be obtained at their output based on the data present at the input in a much faster way than by the traditional direct computation. By embedding such neural networks in an IoT system, we arrive at the concept of AI-IoT, enabling resource-limited devices to perform complex computations simply, quickly and with reduced power consumption.

This Doctoral Thesis describes the use of two sets of acoustic parameters, one to evaluate the psycho-acoustic annoyance in generally large environments and other to evaluate the general acoustic behavior of a room and the intelligibility of speech in it. Due to the complexity of the necessary calculations, CNNs have been used to accelerate them and to be able to perform them inside a WASN node. The CNNs have been designed to minimize the prediction error as much as possible while maintaining a high computational speed. They have been tested with different datasets including real signals acquired with different devices, in order to verify the correct operation in real environments under the influence of different external agents. As a result, an IoT system has been designed and deployed that makes use of neural networks designed for accurate, simple and fast monitoring of acoustic parameters. The AI-IoT system is able to store the data locally or in the cloud and to predict the parameters from unprocessed audio signals (RAW) significantly faster than using direct computation. Adding this to the use of batteries in the nodes and wireless communication, the system allows to analyze a soundscape in a much simpler and faster way than any classical system.

Keywords
Convolutional neural networks, internet of things, psycho-acoustic parameters, regression, room acoustic parameters, soundscape, wireless acoustic sensor networks.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La presente Tesis Doctoral investiga la forma de mejorar las técnicas de monitorización de paisajes sonoros o *soundscapes* extensos y de características acústicas de recintos cerrados, empleando para ello dos conjuntos de parámetros acústicos. Se profundiza también en los algoritmos de cálculo y sistemas de monitorización de los parámetros acústicos buscando la optimización de los mismos y la capacidad de obtenerlos mediante el uso de redes de sensores e inteligencia artificial. Como resultado de esta investigación, se propone un sistema AI-IoT que hace uso del aprendizaje profundo aplicado a redes neuronales convolucionales para calcular un gran abanico de parámetros acústicos que nos proporcionarán un análisis rápido y completo ya sea el caso del paisaje sonoro de un espacio exterior o de una sala a estudiar. Los resultados obtenidos de la investigación han sido publicados en diferentes artículos de revista y comunicaciones de congresos, lo que ha propiciado que esta Tesis se presente como compendio de publicaciones.

En este capítulo, se realiza una introducción a la temática general de esta Tesis, se describen la motivación del trabajo realizado y los objetivos en las secciones 1.1 y 1.2, respectivamente. En la última sección de este capítulo (1.3) se detalla cómo está estructurada la presente memoria.

1.1. Motivación y temática general

La descripción del paisaje sonoro y el comportamiento acústico de diferentes entornos (interiores y exteriores), de la forma en que se realiza hasta la fecha y con los métodos que contempla la normativa actual se puede mejorar sustancialmente gracias a la tecnología actual para proporcionarnos un análisis mucho más completo en cuanto a información proporcionada del entorno a estudiar. Para ello es necesario investigar, desarrollar e implementar procedimientos que hagan un uso conjunto de diferentes tecnologías dispuestas a nuestro alcance, como pueden ser las redes de sensores acústicos inalámbricos (*Wireless Acoustic Sensors Networks, WASN*), la inteligencia artificial y su aplicación al Internet de las Cosas (*Internet of Things, IoT*).

Desde mucho antes de que fuera definido el concepto de *soundscape* por R. Murray Schafer [1], se ha estudiado cómo afecta el sonido que nos rodea a la calidad de vida y a nuestra salud. El concepto de *soundscape* no sólo hace referencia a la totalidad de los sonidos presentes en un área definida, sino a las modificaciones que estos sufren en el tiempo, de manera que un cambio en los mismos representa un cambio estructural es ese espacio, ya sea de la fuente de

sonido o de la física del entorno y nos puede proporcionar información del mismo a muchos niveles. Así, la definición del paisaje sonoro es más compleja de lo que aparenta en un primer momento por lo que se ha descrito con detalle, incluyendo los estudios más habituales a realizar sobre el mismo, en la norma ISO 12913:1 - *Acoustics-Soundscape* [2], ya que amplía el rango de los estudios que evalúan la contaminación acústica presente en una zona de una ciudad o los niveles de sonido percibidos en distintas zonas de un auditorio.

A nivel de estudios de contaminación acústica, en las últimas décadas se han publicado diferentes normas y estándares para hacerle frente, relativos a las magnitudes y métodos de evaluación del ruido ambiental recogidos en la ISO 1996-1 [3] o en la Directiva europea sobre el ruido ambiental (*Environmental Noise Directive, END*) 2002/49/CE [4]. Estas medidas tienen como fondo una gran preocupación por la calidad de vida de los habitantes de los entornos urbanos en términos de molestia acústica subjetiva, pues se ha demostrado que tiene una gran influencia en la salud [5, 6, 7]. Algunas de estas normas se centran en medir los Niveles de Presión Sonora (*Sound Pressure Levels, SPL*) en decibelios (dB) a lo largo del tiempo por lo que no aportan ninguna información de la distribución o contenido frecuencial del sonido, por ejemplo. Además de los niveles SPL, es de interés contar con más parámetros que describan otras características del sonido y tener así, conjuntos de parámetros que en conjunto nos proporcionen una mejor información de cómo es el sonido presente en un soundscape y de cómo son las características acústicas del mismo. En este aspecto se fundamenta la base de esta Tesis, recogiendo dos conjuntos de parámetros: en un primer caso para analizar la molestia psico-acústica que provocan los sonidos presentes en un paisaje sonoro y en un segundo caso para analizar el comportamiento acústico de una sala, pues este afectará a los sonidos producidos en la misma.

Si deseamos tener un análisis preciso de las características del sonido presente en un soundscape o de las características acústicas del espacio que lo comprende de forma detallada, es esencial tomar un gran número de muestras en diferentes puntos, lo que nos aportará información de la distribución espacial a parte de la temporal de los parámetros acústicos analizados, como se puede observar en la Figura 1.1. En la parte izquierda de la Figura 1.1 se pueden ver destacados en azul diferentes puntos de interés a monitorizar a la hora de estudiar el paisaje sonoro de un barrio urbano, y analizar cómo es el sonido presente en zonas de ocio, jardines, o residenciales y obtener un análisis espacial detallado. En la parte derecha de la Figura 1.1 se puede ver una sala destinada por ejemplo a la docencia o a la realización de conferencias. En azul se destacan diferentes puntos donde sería interesante analizar el comportamiento acústico para detectar si hay zonas que presentan problemas debidos a la distancia, reverberación u otros efectos en zonas periféricas y centrales, para poder actuar en consecuencia. Es fácil deducir que desplegar un sistema de sensores cableado presenta numerosas dificultades en el caso de una sala y más aún en el caso del paisaje sonoro de un barrio o una ciudad completa.

Las redes inalámbricas de sensores facilitan esta tarea pues permiten monitorizar en numerosos puntos de forma fácil y rápida. En concreto las WASN, con sensores específicamente destinados al audio, están incrementando su popularidad en el campo del procesado de señales acústicas y están siendo empleadas como solución innovadora a problemas clásicos como pueden ser la localización de fuentes acústicas [8, 9], la detección y clasificación de eventos acústicos [10], la evaluación del ruido ambiental [11] o incluso el campo de los asistentes domésticos [12]. En este apartado, esta Tesis describe cómo se ha afrontado la monitorización de los diferentes parámetros acústicos empleando una WASN para permitir un muestreo distribuido de los mismos. Los nodos de la WASN están conectados a internet y pueden ser gestionados de forma local o remota, implementando un sistema IoT que permite tanto la monitorización continua, como el almacenamiento y consulta de los datos también de forma local y remota, así como su instalación de forma permanente o móvil.

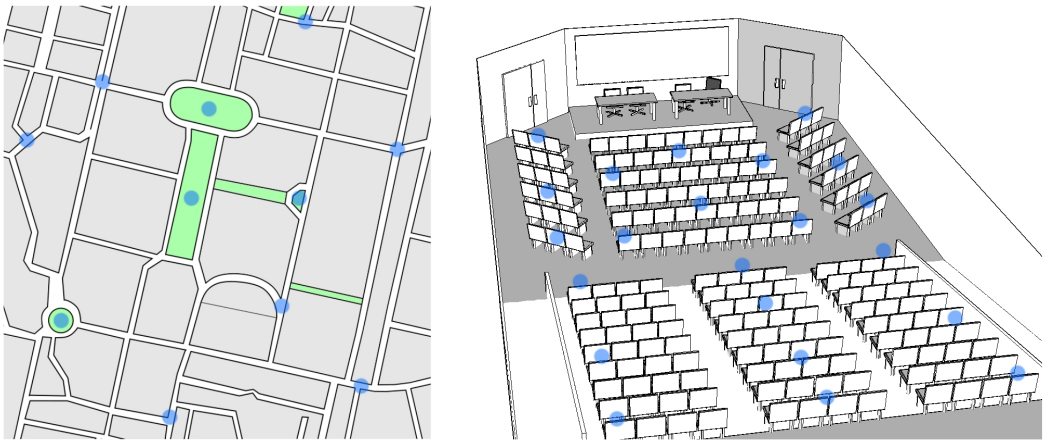


Figura 1.1: Puntos de monitorización acústica de interés, en una ciudad (izquierda) y en una sala (derecha).

En la mayoría de casos, los nodos de los sistemas IoT o las WASN están formados por ordenadores de placa única o *Single Board Computers (SBC)*, que engloban en el mismo circuito impreso el microprocesador, la memoria RAM, el almacenamiento no volátil y diferentes puertos de entrada y salida (E/S). Algunos ejemplos de estos dispositivos son la placa Raspberry Pi [13], la Jetson Nano de Nvidia [14] o la Thinker Board de Asus [15]. En algunos casos incluyen además interfaces de conectividad inalámbrica o cableada, de sonido, o unidades de procesado gráfico (GPU). Con esto se consigue tener un dispositivo compacto que incorpora la mayoría de las funcionalidades necesarias para las aplicaciones más comunes de estos dispositivos. Para aplicaciones específicas las conexiones de E/S suelen incluir puertos serie USB y permitir diferentes protocolos de comunicación ($I2C$, SPI, etc), además de puertos digitales y analógicos, lo que permite conectar una gran variedad de sensores específicos disponibles en el mercado. La reducción del tamaño, del consumo y un coste económico contenido se suman a las ventajas de estos dispositivos SBC, sin embargo estas características a menudo limitan su capacidad de procesamiento por lo que se debe prestar especial atención en optimizar al máximo el código a ejecutar en los mismos. En nuestro caso, el cálculo de los parámetros acústicos requiere de algoritmos complejos que realizan muchas operaciones de filtrado en diversas bandas de frecuencia, por lo que su coste computacional puede ser elevado. Para los parámetros relacionados con la molestia psico-acústica, es interesante disponer de un sistema de monitorización en tiempo real o lo más aproximado posible, ya que se analiza el audio que se está percibiendo en un instante y en la mayoría de los casos se desean ver los valores de los parámetros en el momento en que se producen ciertos sonidos. En el caso de los parámetros de sala, estos ayudan a describir el comportamiento acústico de la sala, que no debería cambiar si no cambian la geometría o los materiales que la forman, sin embargo, es muy útil también tener los valores de los parámetros en el tiempo más breve posible para realizar otro muestreo en posiciones diferentes por ejemplo. Si empleamos menos tiempo en el cálculo de los parámetros reducimos el gasto energético, permitiendo un incremento de temperatura menor en el sistema y un ahorro de batería en el caso de un sistema IoT alimentado de esta manera.

La inteligencia artificial, en forma de redes neuronales está ganando protagonismo en el campo del procesado de señal acústica en los últimos años en tareas como la localización de fuentes [16, 17], la detección de eventos acústicos [18, 19], la separación y clasificación de fuentes sonoras tanto de forma general [20, 21, 22] como de ámbitos concretos, como puede ser la medicina [23]. Sin embargo, no está tan extendido el uso de redes neuronales en el cálculo de parámetros acústicos, donde pueden ser de gran ayuda al permitir reducir el tiempo de cálculo de los mismos de forma considerable. En esta Tesis Doctoral se describe

el diseño e implementación de dos redes neuronales convolucionales (CNN), que permiten la predicción de los parámetros relativos a la molestia psico-acústica y de un conjunto de parámetros acústicos de sala con muy buena precisión. Estas redes neuronales se han incluido en el sistema IoT diseñado permitiendo su funcionamiento en tiempo real en le caso de monitorización continua, gracias a su elevada velocidad de cálculo.

En el siguiente punto se enumeran los objetivos planteados en la presente Tesis.

1.2. Objetivos

Esta Tesis Doctoral tiene como objetivo principal la investigación y desarrollo de nuevas herramientas para el estudio acústico más completo de los paisajes sonoros y del comportamiento acústico de diferentes recintos. Para lograrlo se emplean dos conjuntos de parámetros acústicos y se hace uso de los últimos avances en materia de inteligencia artificial y redes de sensores para obtener los parámetros a partir de señales de audio sin procesar. Este objetivo principal puede ser dividido en los siguientes sub-objetivos:

- Revisar, testear y evaluar métodos, algoritmos y estándares usados en la monitorización de soundscapes y caracterización acústica de espacios para tener una visión general de las soluciones más avanzadas a los problemas abordados en la Tesis. Hacer especial hincapié en los parámetros de molestia psico-acústica y en los parámetros de sala para lograr un estudio más preciso y completo.

- Desarrollar sistemas que permitan la monitorización distribuida de diversos parámetros acústicos mediante el uso de redes de sensores. Obtener así una evaluación acústica objetiva del entorno a estudiar más allá de lo que establece la normativa vigente.

- Desarrollar herramientas integradas en los nodos de la red de sensores para permitir la realización de los cálculos pertinentes en los propios nodos (*Edge Computing*) y el funcionamiento en tiempo real evitando la sobrecarga de la red y de los servidores enviando el mínimo volumen de datos posible.

- Desarrollar herramientas que permitan el control, sincronización, recolección y almacenamiento de la información calculada por los nodos de la red de sensores.

- Realizar pruebas de campo de la tecnología y sistemas desarrollados en entornos de diversos tamaños, tanto virtuales como reales. Probar los sistemas desarrollados tanto en entornos exteriores o de tránsito elevado como en salas de dimensiones más reducidas durante la realización de las actividades comunes a cada zona.

1.3. Estructura de la Tesis Doctoral

La presente Tesis Doctoral ha sido financiada mediante la Ayuda para contratos predoctorales BES-2017-082340, financiada por el Ministerio de Ciencia e Innovación, la Agencia Estatal de Investigación, (MCIN/AEI /10.13039/501100011033) y por "FSE invierte en tu futuro". Esta ayuda está asociada al proyecto de investigación BIA2016-76957-C3-1-R, "Urbauramon, Herramientas inteligentes para la gestión y control del paisaje sonoro urbano. Definición de protocolos de monitorización y auralización. Intervención en el Patrimonio Sonoro", correspondiente al Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad, en el Marco del Plan Estatal de Investigación Científica y Técnica

y de Innovación 2013-2016.

Esta Tesis Doctoral se presenta como compendio de publicaciones, según la normativa de la Escuela de Doctorado de la Universitat de València y su Programa de Doctorado en Tecnologías de la Información, Comunicaciones y Computación del Departamento de Informática. Siguiendo los requisitos establecidos, se ha estructurado el presente documento en cinco capítulos que se detallan a continuación.

En el Capítulo 1 se aborda la temática general de la tesis que plantea el problema a investigar, los objetivos de esta investigación y la estructura de este documento. A continuación, en el Capítulo 2 se profundiza en aspectos del estado del arte relacionados con la parametrización y monitorización acústica, necesarios para encuadrar mejor el resto de contenidos del capítulo y los siguientes. Se proporciona una descripción profunda de los algoritmos de cálculo directo de parámetros acústicos y la evaluación de los mismos realizada. Asimismo se describe también el sistema de monitorización IoT diseñado y la representación de resultados analíticos que permite.

En el Capítulo 3 presenta conceptos básicos de las redes neuronales empleados más adelante en el capítulo para presentar el núcleo de esta Tesis, el diseño de 2 redes neuronales convolucionales que mediante aprendizaje profundo permiten predecir los parámetros acústicos de cada conjunto definido. Junto a una descripción profunda de las bases de datos y configuración empleada para el entrenamiento de los modelos, se detallan también las numerosas pruebas d rendimiento realizadas, así como los resultados de su incorporación al sistema IoT implementado.

Seguidamente, en el Capítulo 4 se describen las contribuciones más importantes de esta investigación, así como las diferentes publicaciones derivadas de estas. Posteriormente, en el Capítulo 5 se exponen las conclusiones de esta Tesis Doctoral junto con algunas ideas sobre posibles trabajos futuros.

Para finalizar, siguiendo las indicaciones de la normativa se incluyen los Anexos donde se puede consultar la versión publicada en revistas (incluidas en el índice JCR), de los 4 artículos incluidos en el compendio:

- **Anexo A:** “Computation of psycho-acoustic annoyance using deep neural networks”, Applied Sciences-Basel, 9 (15), 1-12, 2019, [24].
- **Anexo B:** “Enabling real-time computation of psycho-acoustic parameters in acoustic sensors using convolutional neural networks”, IEEE - Sensors Journal, 20 (19), 11429-11438, 2020, [25].
- **Anexo C:** “Speech Intelligibility Analysis and Approximation to Room Parameters through the Internet of Things”, Applied Sciences-Basel, 11 (4), 1430-1440, 2021, [26].
- **Anexo D** “AI-IoT Platform for Blind Estimation of Room Acoustic Parameters Based on Deep Neural Networks”, IEEE - Internet Of Things Journal, 2022, [27].

Capítulo 2

Monitorización Acústica

En este capítulo se profundiza en el problema planteado acerca de la mejora de la monitorización y análisis acústicos. Inicialmente se aborda la descripción de los parámetros acústicos y cómo se emplean en los actuales sistemas de análisis, que servirá de base para seleccionar dos conjuntos de parámetros acústicos para implementar análisis. El primer conjunto de parámetros está orientado a medir la molestia psico-acústica producida por los sonidos presentes en un instante en el entorno estudiado. El segundo conjunto de parámetros acústicos está orientado a medir el comportamiento acústico característico de una sala además de su influencia en la transmisión del habla.

A continuación se describe el estado del arte en lo concerniente a sistemas de medición para obtener los parámetros psicoacústicos y los parámetros acústicos de sala, presentando el sistema cableado empleado en esta tesis, así como el sistema de monitorización IoT implementado junto con las posibilidades de análisis que permite. De la misma manera se estudian los algoritmos de cálculo directo de parámetros y se detallan las diversas pruebas de rendimiento realizadas.

2.1. Parámetros Acústicos

La contaminación acústica ha despertando mucho interés en los últimos tiempos al estudiarse su impacto sobre la salud humana, ya que está relacionada con afecciones tan distantes como enfermedades cardiovasculares, problemas de cognición en niños o trastornos psicológicos [28]. Aunque la contaminación acústica está directamente relacionada con la actividad humana, en diversos estudios se demuestra que afecta también a la fauna tanto de entornos terrestres [29, 30] como marinos [31], por lo que es de especial interés la implementación de sistemas que permitan monitorizar y analizar *soundscapes*. En entornos urbanos y sus proximidades es un problema de importancia al concentrarse la mayor parte de la actividad humana, por lo que la Comisión Europea (CE) consideró la adopción de una directiva internacional, la Directiva sobre el ruido ambiental (END 2002/49/CE), en 2002. En este documento, la CE exige que se realicen mediciones en las ciudades de más de 250.000 habitantes, para recoger información sobre la exposición al ruido con el fin de elaborar planes de acción locales y proporcionar mapas de niveles de contaminación acústica. Los estudios a realizar sobre un *soundscape* se encuentran definidos en la norma ISO 12913-1 [2], así como las magnitudes y métodos a emplear para la evaluación del ruido ambiental están definidos en la ISO 1996-1 [3], ambas revisadas en el año 2020 y que sirven como guía prácticamente para la totalidad de estudios que se realizan. Algunos ejemplos de estos estudios detallados son

los mapas de control del ruido desarrollados en Pekín [32] y Londres [33] destinados describir el ruido ambiental o el proyecto *Smart Santander* [34], mediante el que se ha desplegado una red de sensores de ruido y de parámetros medioambientales (temperatura, CO_2, luz o irrigación), como se puede ver en la Figura 2.1.

Figura 2.1: Imagen de un mapa de ruido de la zona centro de Santander del proyecto Smart Santander.

Estos estudios se basan principalmente en medir los niveles de presión sonora SPL (*Sound Pressure Level*) en un instante dado o en medir la sonoridad o volumen equivalente, L_eq (*Loudness Equivalent*) en un intervalo de tiempo. Esta magnitud no contempla aspectos espacio-temporales ni frecuenciales de la señal, por lo que no describe con precisión la molestia acústica producida por ciertos sonidos y puede resultar insuficiente para los estudios de precisión que son necesarios, como veremos más adelante. Por lo tanto, es necesario contar con más parámetros para que en su conjunto nos proporcionen una descripción más precisa del *soundscape* a evaluar y analizar cómo afecta el sonido presente a las personas que lo habitan. De estudiar mejor la percepción del sonido por los humanos y la respuesta de carácter psicológico que origina se encarga la psicoacústica. Numerosos expertos se han dedicado a su estudio, pero los más relevantes como son Hugo Fastl, Eberhard Zwicker [35] o Brian C.J. Moore [36] nos han llevado a seleccionar el modelo de molestia psicoacústica de Zwicker y los 4 parámetros que lo componen, junto con un indicador general de molestia, que se describen en la Sección 2.1.1 y que se emplearán para desarrollar el sistema de monitorización acústica que recoge esta Tesis.

Parámetros como la sonoridad equivalente (L_{eq}) por ejemplo, así como el Loudness (N) y el resto de parámetros psicoacústicos recogidos por el modelo de Zwicker, se pueden emplear para analizar el sonido presente en una zona y que conforma el paisaje sonoro a estudiar. Nos dan información del sonido presente en cada momento en un entorno y se pueden registrar a lo largo del tiempo para su estudio, sin embargo, no nos aportan información acerca de cómo se comporta acústicamente el espacio estudiado, independientemente del sonido presente.

Si en vez de analizar un barrio o una zona de ocio de una ciudad, deseamos evaluar las características acústicas de un entorno más reducido o cerrado como una sala, debemos medir otros parámetros acústicos que nos den información objetiva de cómo se va a comportar el sonido en esa sala. Estos se denominan *parámetros acústicos de sala* y complementan a los parámetros psicoacústicos mencionados hasta ahora, pues evalúan las características de una sala más allá del sonido presente en el momento en la misma y están más orientados a recintos de dimensiones más contenidas.

Los parámetros acústicos de sala nos proporcionan una descripción precisa y objetiva del comportamiento acústico de un recinto, sea cual sea su naturaleza y su uso, desde un entorno industrial a un auditorio, por lo que son de gran interés para controlar la percepción del sonido en estos emplazamientos. Comenzando con los primeros estudios de principios del

siglo XX publicados por W.C. Sabine [37] sobre resonancia, reverberación y sonoridad, se ha desarrollado una profunda investigación sobre diferentes parámetros acústicos de sala que tratan de medir los efectos acústicos producidos por las características arquitectónicas de los recintos, para mediante su estudio conseguir mejorar los negativos y realzar los deseados en cada caso. Además de un abanico de parámetros acústicos de sala como pueden ser la reverberación (RT de *Reverberation Time*), la claridad (C), la definición (D) o el brillo (Br), los métodos de medida crecían en variedad hasta la aparición de la norma ISO 3382: *Acoustics — Measurement of room acoustic parameters, Parts 1, 2 and 3* [38, 39, 40] que se encarga de estandarizar un conjunto de parámetros acústicos y los métodos de medida de los mismos para sentar las bases de estudios acústicos estándar. Esta norma contempla sobretodo parámetros acústicos relacionados con la reverberación y durante muchos años la primera parte se dedicó a los métodos de medida del tiempo de reverberación y otros parámetros acústicos en espacios de actuación y la segunda parte dedicada a salas ordinarias. Además de RT y el resto de parámetros acústicos recogidos en la norma ISO 3382, relacionados como decíamos con la reverberación y la distribución energética de la señal acústica, existen otros orientados a analizar la transmisión oral de información en salas como pueden ser el Indice de Transmisión del Habla o *Speech Transmission Index*, (STI) o el Indice de Inteligibilidad del Habla o SII (*Speech Inteligibility Index*).

Por lo tanto, al existir un número considerable de parámetros acústicos de sala orientados a distintos tipos de análisis, en nuestro caso, tal y como se hace en otros estudios de investigación, hemos escogido un conjunto de 5 parámetros acústicos de sala que nos proporcionan una descripción de la reverberación, la distribución energética del sonido y de la inteligibilidad del habla de la sala a evaluar. Si bien es verdad, el conjunto está balanceado ligeramente hacia parámetros relacionados con la voz y el análisis del comportamiento acústico de salas en este ámbito, ya que deseamos orientarlo al análisis de espacios docentes, como se describe en la Sección 2.1.2.

2.1.1. Parámetros de molestia psicoacústica y modelo de Zwicker

Una de las magnitudes más básicas para medir la intensidad del sonido es el nivel SPL (Ecuación 2.1), que nos proporciona el valor de presión sonora en cada instante de muestreo en decibelios (dB), dividiendo la presión sonora eficaz por una de referencia, pero este valor es continuo y no es útil para estudiar intervalos largos de tiempo, pues ciertos valores puntuales pueden conducir a error, falseando el promediado o pasando desapercibidos pese a ser importantes. Por lo que en un primer intento de optimización se emplea el valor de presión sonora equivalente L_{eq}. El nivel L_{eq}, definido en la Ecuación 2.2, nos da el nivel equivalente de energía en decibelios (dB) que produciría esa señal en un intervalo de tiempo.

$$SPL = 20log[\frac{p_a}{p_0}] \tag{2.1}$$

$$L_{eq} = 10log[\frac{1}{(t_2 - t_1)} \int_{t_1}^{t_2} \frac{p_a^2}{p_0^2} \mathrm{d}x] \tag{2.2}$$

En la Ecuación 2.2, podemos ver que se divide la presión sonora eficaz o adquirida en cada instante p_A, por una presión de referencia p_0, (típicamente de 20 µPa como en el caso de SPL), sin más consideraciones que la presión sonora, por ejemplo respecto a contenido espectral, o a efectos espacio-temporales del sonido monitorizado, como se puede ver de forma gráfica en la Figura 2.2.

Como todo el cálculo se realiza en el dominio del tiempo, no se representa ninguna banda específica de frecuencia, por lo que se ve claramente que esta magnitud por sí sola no aporta suficiente información. Para solucionar parcialmente esto sin emplear parámetros distintos se incorpora el cálculo de L_{eq} con ponderación A o $A - Weighted$, pasándose a denominar

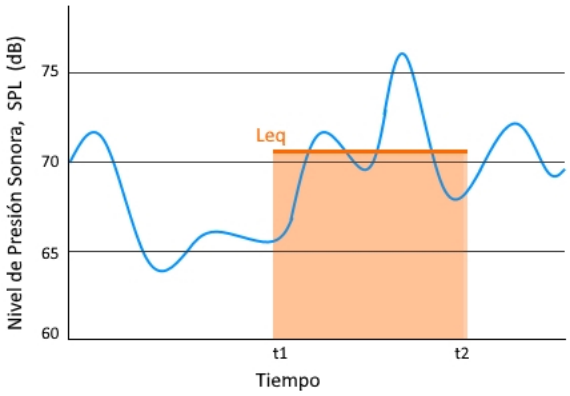
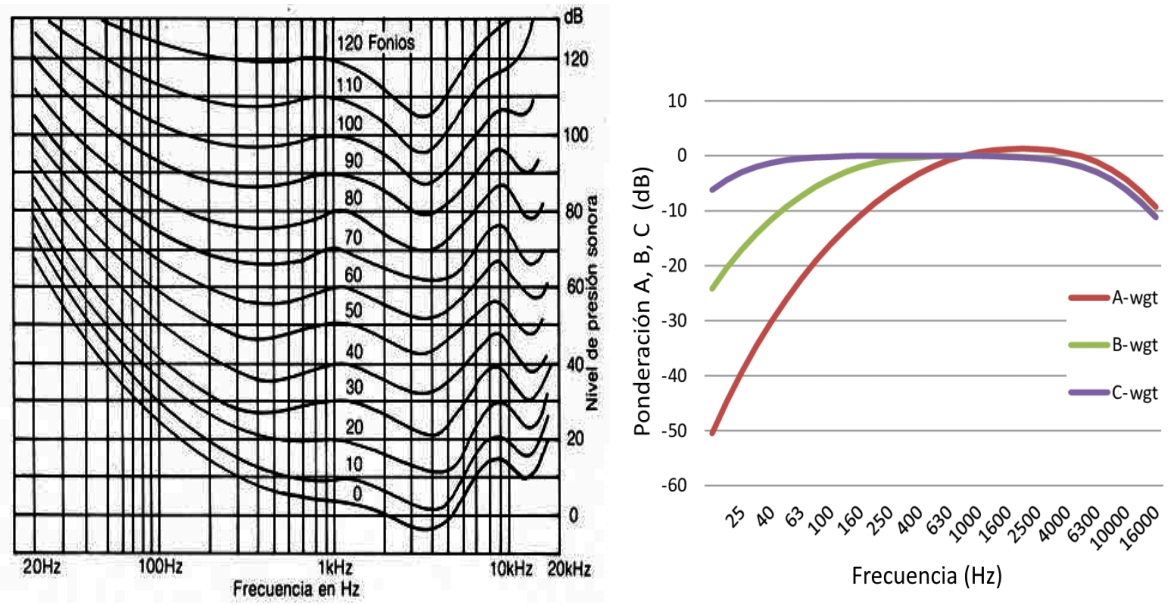


Figura 2.2: Medida de SPL a lo largo del tiempo y nivel equivalente (L_{eq}) en un intervalo.

LA_{eq} (medido en dB(A)). Aunque hay otras ponderaciones, en estudios oficiales se emplea la ponderación A, que está contemplada por la norma ISO 226:2003 [41] y que emplea unos coeficientes que se pueden ver en la Figura 2.3a para ponderar los valores de L_{eq} calculados y que estos se adapten ligeramente al patrón de percepción frecuencial que tiene el oído humano que se puede observar también en la Figura 2.3b.



(a) Contornos característicos de igual sonoridad.

(b) Valores de ponderación A, B y C.

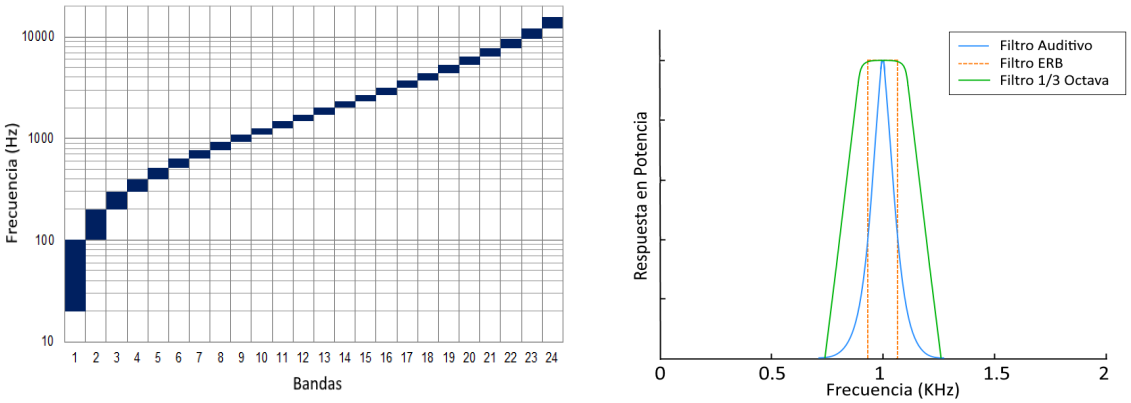
Figura 2.3: Patrón de percepción acústica del oído humano y valores de ponderación según ISO 226:2003.

Así, el nivel de sonoridad ponderado LA_{eq} incorpora algo de información espectral a la medida, se encuentra respaldado por la normativa vigente y se emplea en la práctica totalidad de aplicaciones de medida de ruido ambiental. No obstante, aunque es mejor que el simple valor de nivel de presión sonora SPL, no es una magnitud completa o perfecta para el estudio de cómo afecta el ruido ambiente a los humanos, como demuestran diversos estudios [42, 43, 44]. Esto es debido a que no contempla numerosos efectos, como por ejemplo la duración de los sonidos (sonidos de más de 200 ms son más molestos), la forma en que se miden (campo abierto o salas cerradas) o si nuestro patrón de audición difiere mucho del patrón medio contemplado en la ISO, ya que diferencias superiores a 5 dB suelen ser habituales y sonidos de bajas frecuencias llegar a ser muy molestos para algunas personas, siendo supri-

mido por los equipos de medida habituales con ponderación A.

Los parámetros psicoacústicos permiten cuantificar el grado de molestia subjetiva que ciertos sonidos producen en las personas de forma totalmente objetiva. De los estudios de los expertos Hugo Fastl y Eberhard Zwicker se ha seleccionado el modelo de molestia de Zwicker [35] que se basa en 4 parámetros psicoacústicos: Sonoridad o *Loudness* (N), Agudeza o *Sharpness* (S), Rugosidad o *Roughness* (R) y Fuerza de fluctuación o *Fluctuation Strength* (F). Gracias a esos 4 parámetros psicoacústicos el modelo de Zwicker permite calcular un marcador de molestia general PA (*Psicoacoustic Annoyance*) a modo de resumen del resto de parámetros, para expresar la molestia de un sonido con un solo valor. Este modelo se basa en la anatomía del oído humano, que realiza un análisis en frecuencia del sonido percibido mediante un comportamiento que podríamos comparar con un banco de filtros con un ancho de banda dependiente de la frecuencia y que están implementados en la membrana basilar situada en la cóclea, órgano sensorial de la audición situada en el oído interno.

Estos filtros agrupan los estímulos sonoros que tienen un contenido frecuencial cercano en una misma banda, que se llama entonces banda crítica. Debido a este comportamiento, los parámetros psicoacústicos se analizan empleando bandas críticas [45] cuyo ancho de banda es equivalente a los implementados por el oído. Para estas bandas críticas se definen dos escalas psicoacústicas, la escala Bark propuesta por Zwicker [46] que se puede ver en la Figura 2.4a y la escala ERB o de Banda Rectangular Equivalente (*Equivalent Rectangular Bandwidth*), propuesta por Moore y Glasberg [47] e implementa una revisión de la escala Bark modelando los filtros como rectangulares pasa-banda, como se puede ver en la Figura 2.4b. La escala psicoacústica Bark cuenta con 24 bandas críticas y se mide en Barks, en la escala ERB (medida en ERBs), en cambio las bandas son más estrechas y cuenta con 33 bandas críticas. Aunque el cálculo teórico de los parámetros acústicos se define normalmente para escalas Bark o ERB, en la práctica se emplea una solución de compromiso intermedia que permite implementar los filtros por banda de una forma más sencilla, se trata de la escala en tercios de octava. En la Figura 2.4b se puede ver una comparación entre las respuestas en frecuencia del filtro auditivo real, el ERB y el de 1/3 de octava. Las escala en 1/3 de octava cuenta con 31 bandas críticas y las características de los filtros están definidas en la norma UNE-EN 61260 [48].



(a) 24 Bandas críticas de la escala Bark.

(b) Ejemplo de filtro ERB para 1 KHz.

Figura 2.4: Banda críticas en la escala Bark, ERB y 1/3 de Octava

Estas escalas se emplean para calcular los parámetros psicoacústicos del modelo de molestia de Zwicker, *Loudness*, *Sharpness*, *Roughness* y *Fluctuation Strength*, que se definen de manera esquemática a continuación.

Loudness (N) mide la sensación humana de intensidad de sonido o sonoridad, se representa como N y se mide en *sones* empleando una escala lineal, como explicaremos a continuación. Cabe mencionar, que si se representa como L, se mide en *phones* y se emplea una escala logarítmica, como es el caso de L_{eq} o LA_{eq}, aunque son escalas menos realistas. El método de cálculo fue estandarizado inicialmente por Zwicker en la norma ISO 532B [49] y se revisó para sonidos variantes en el tiempo e incluyendo código informático para acelerar los cálculos en la norma DIN 45631/A1 [50]. Es esta la norma la que hemos seguido para el cálculo de N debido a que contempla mejor el patrón de percepción del oído humano y a que es más adecuando para sonidos poco estacionarios, pues L_{eq} induciría error, como se puede ver fácilmente en la Figura 2.2 si pensamos en una señal que presente picos de SPL. El cálculo que podemos ver en la Ecuación 2.3 se basa en sumar el loudness específico de cada banda crítica $N'(z)$ en la escala Bark, ponderado por el ancho de banda de cada banda crítica.

$$N = \sum_{z=0}^{24Bark} N'(z) \cdot \Delta_z \tag{2.3}$$

En nuestro algoritmo para el cálculo de N, el audio de entrada se divide en ventanas de 1 segundo empleando una ventana Hann y solape del 50 % como máximo si se desea más precisión de cálculo o sin emplear solape en las ventanas si se desea más velocidad de cálculo. Siguiendo el método propuesto en la norma DIN 45631, se ha empleado un banco de filtros en 1/3 de octava como equivalentes a la escala Bark a la hora de calcular los valores específicos de Loudness por banda $N'(z)$ a partir de los niveles SPL del audio de entrada.

Sharpness (S) mide la cantidad de componentes de alta frecuencia respecto al ancho de banda total que posee un sonido. A nivel de molestia psicoacústica, los sonidos que poseen un contenido elevado de alta frecuencia se presentan como muy agudos y molestos. El cálculo de S tal y como contempla el modelo de Zwicker está estandarizado en la norma DIN 45692 [51], se mide en *Acums* en una escala lineal y se calcula mediante la Ecuación 2.4.

$$S = C_S \cdot \frac{\sum_{z=0}^{24Bark} N'(z) \cdot g_s(z) \cdot z\Delta_z}{N}. \tag{2.4}$$

Para el cálculo de S se emplean los valores antes calculados de Loudness específico por bandas de 1/3 de octava $N'(z)$, ponderados por la función $g_s(z)$, que incrementa los valores de ponderación a medida que aumenta el número de banda crítica dando así más peso a las componentes de frecuencia alta del sonido, como se puede ver en la Figura 2.5 . C_S es una constante de calibración para obtener un S de 1 Acum con un tono de 1 kHz y 60 dB SPL. S junto a N son parámetros relacionados con la distribución energética de la señal a lo largo de su espectro, uno orientado a las bandas de alta frecuencia y el otro orientado a las bandas donde el oído humano es más sensible.

Roughness (R) mide la aspereza o rugosidad del sonido, es decir, las variaciones rápidas del mismo, en amplitud y en frecuencia incluso cuando se mantienen N o L_{eq} constantes. En términos perceptivos la fluctuación rápida de un sonido puede parecer constante pero genera rugosidad molesta en el mismo, como por ejemplo el sonido resultante al pronunciar la letra r de forma contínua o el emitido por un motor de combustión girando a bajas revoluciones. Estas modulaciones del sonido comienzan a escucharse como un sonido contínuo desde los 15 Hz, alcanzando su intensidad máxima en los 70 Hz y desapareciendo en torno a los 300 Hz. R se mide en la unidad perceptual *Asper*

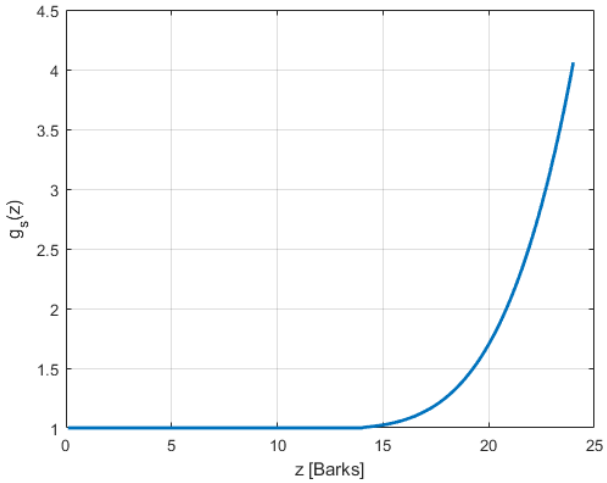


Figura 2.5: Función $g_s(z)$ de ponderación empleada en el cálculo del Sharpness.

y analiza el efecto de modular en frecuencias cercanas a los 70 Hz cada banda crítica del sonido analizado. En el caso de R no se ha estandarizado su cálculo todavía y se han creado varios modelos similares basados en el parámetro definido por Helmholtz y empleado por Zwicker en su modelo de molestia psicoacústica. Algunos de los más destacados son los modelos de R. Sottek y J. Becker [52], la optimización de P. Daniel y R. Webber [53] o su modelado computacional realizada por V. Jourdes [54] que hemos implementado en nuestro caso. Para el cálculo de R (Ecuación 2.5) se emplea la escala ERB donde para cada ERB i se calcula una función de ponderación de la banda $g_r(z_i)$ por la profundidad de modulación de la misma m y la correlación cruzada k entre las envolventes de las bandas ERB i e $i-2$.

$$R = C_R \sum_{i=0,5}^{33\,ERB} \left(g_r(z_i) \cdot m_i \cdot k_{i-2} \cdot k_i\right)^2 \tag{2.5}$$

C_R es una constante de calibración para obtener 1 Asper si medimos una señal portadora de 1 kHz modulada al 100 % en amplitud por una señal de 70 Hz. Para esta señal portadora de calibración de 1 kHz podemos ver en la Figura 2.6 cómo variará R en función de la profundidad de modulación en amplitud de la misma al emplear una frecuencia moduladora de 70 Hz. En esta Figura 2.6, el punto en la esquina superior derecha indica el sonido estándar, que produce la rugosidad de 1 Asper y la línea discontinua indica una aproximación lineal útil.

Esquemáticamente el algoritmo de cálculo compara por bandas ERB si la señal a analizar correla con señales moduladas a distinta profundidad y nos da un valor de R que será más elevado conforme haya más bandas que presenten modulaciones elevadas. La señal de entrada se muestrea y se divide en fragmentos de 1 segundo de duración, como en los casos anteriores, pero esta vez se emplea una ventana Blackmann sin superposición o con un solape del 50 % según el caso, si se desea más velocidad de cálculo o más precisión.

Fluctuation Strength (F) mide la percepción subjetiva de fluctuaciones en la señal de audio producida por modulaciones lentas en la amplitud del mismo. Si R está centrado en las fluctuaciones a alta frecuencias, Fluctuation strength se centra en las modulaciones a muy bajas frecuencias, pues este efecto es más notable a 4 Hz y desaparece a los 20 Hz. El sonido modulado de una sirena de ambulancia nos puede dar una idea

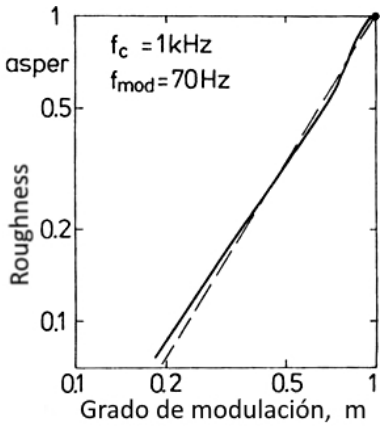


Figura 2.6: R para tono de 1 kHz en función del grado de modulación en frecuencia con moduladora de 70 Hz.

de las modulaciones a las que nos referimos. Como en el caso de R, su cálculo no está estandarizado pero el descrito por Zwicker es similar al de R con la diferencia de las frecuencias de modulación empleadas en cada banda ERB, que estarán centradas en 4 Hz, como se puede ver en la Ecuación 2.6. La unidad de medida de F es el $Vacil$. En el calculo de F para cada ERB i se calcula una función de ponderación de la banda $g_f(z_i)$ por la profundidad de modulación m y la correlación cruzada k entre las envolventes de las bandas ERB i e $i-2$. C_F es una constante de calibración que permite obtener 1 vacil al analizar una señal de 1 kHz modulada al 100 % en amplitud con una señal de 4 Hz.

$$F = C_F \sum_{i=0,5}^{33\,ERB} \left(g_f(z_i) \cdot m_i \cdot k_{i-2} \cdot k_i\right)^2 . \tag{2.6}$$

Como en el caso de R, esquemáticamente nuestro algoritmo de cálculo compara por bandas ERB si la señal a analizar correla con señales moduladas a distinta profundidad en torno a los 4 Hz y nos da un valor de F que refleja las modulaciones en baja frecuencia presentes en la señal a analizar. Como en el caso anterior, para el cálculo de F la señal de entrada se muestrea y se divide en fragmentos de 1 segundo de duración y se emplea una ventana Blackmann sin solape o con un solape del 50 % dependiendo de si prima la precisión o la velocidad en el análisis.

El parámetro F ha sido el responsable principal de que hayamos escogido una ventana de sonido de 1 segundo de duración. En la mayoría de estudios mencionados para el cálculo de N, S o incluso R se emplean generalmente ventanas de 200 ms a 500 ms de duración pues permiten capturar bien los efectos estudiados por esos parámetros, además al emplear ventanas de tiempo más pequeñas las señales a analizar contienen menos muestras y los cálculos se pueden realizar en menos tiempo. Sin embargo, en el caso de F, en las diferentes pruebas realizadas al evaluarse modulaciones de muy baja frecuencia de la señal, si no empleábamos ventanas de más de 800 ms ciertos sonidos con modulaciones molestas presentes, podían pasar "desapercibidosz no reflejarse en el valor de F calculado de los mismos. Para garantizar el cálculo correcto de todos los parámetros del modelo de molestia de Zwicker definimos por lo tanto una ventana de tiempo de 1 segundo de duración.

Psycho-acoustic Annoyance (PA) o Molestia Psicoacústica es el indicador general que resume el modelo de molestia psicoacústica de Zwicker. Describe cuantitativamente el nivel de molestia sonora dadas las características físicas de la señal de audio analizada,

basándose en la ponderación de los valores calculados de N, S, R y F. Una vez obtenidos los parámetros anteriores, podemos calcular PA mediante la Ecuación 2.7a y obtener un indicador objetivo de molestia que sintetiza la aportación de los 4 parámetros psicoacústicos mediante valores de ponderación de cada uno. El valor de PA no se acota en los cálculos, pero como norma general está en el intervalo [0, 100] por lo que podemos asimilarlo al porcentaje de molestia psicoacústica de un sonido, con 0 % de molestia para los sonidos nada molestos un 100 % para los extremadamente molestos, simplificando así su evaluación.

$$PA = N\left(1 + \sqrt{{w_S}^2 + {w_{FR}}^2}\right), \tag{2.7a}$$

$$w_S = \frac{(S - 1{,}75) \cdot \log(N + 10)}{4}, \tag{2.7b}$$

$$w_{FR} = \frac{2{,}18 \cdot (0{,}4F + 0{,}6R)}{N^{0{,}4}}. \tag{2.7c}$$

Para un análisis exhaustivo de las características molestas de un sonido deberemos evaluar los parámetros calculados individualmente, sin embargo PA es sumamente útil para tener una idea general de la molestia psicoacústica del sonido monitorizado observando un solo parámetro. Aunque el indicador general de molestia psicoacústica PA no se define como un parámetro psicoacústico en sí, al contener información del resto de parámetros posee una importancia notable y lo hemos empleado como un parámetro más, por lo que a lo largo de esta Tesis, como en otros estudios, cuando hacemos referencia al conjunto de parámetros psicoacústicos seleccionado, se incluyen los 4 parámetros psicoacústicos reales: N, S, R y F junto con PA, aunque en algunos casos no lo nombremos explícitamente como indicador general de molestia.

En la Figura 2.7 se puede ver esquemáticamente el algoritmo diseñado para implementar el modelo de molestia psicoacústica de Zwicker al completo. Como hemos descrito anteriormente, el audio de entrada se divide en ventanas de 1 segundo de duración y se multiplica por una función Hann para proceder al cálculo de N y S y por una Blackman para el caso de R y F. El solape de las ventanas puede ser del 50 % si se desea un análisis más exhaustivo o inexistente si se desea más velocidad, buscando una primera optimización de los tiempos de cálculo. La calibración se ha llevado a cabo siguiendo instrucciones de la normativa pertinente y se ha verificado mediante el software *ArtemiS Suite*[1] .

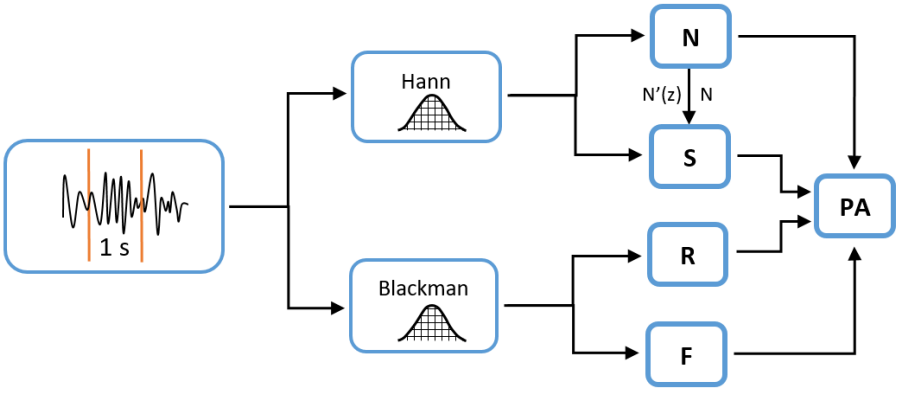


Figura 2.7: Algoritmo de cálculo del modelo Zwicker.

[1] https://www.head-acoustics.com/products/analysis-software/artemis-suite

Los parámetros N y S son los más rápidos de calcular, pues sus algoritmos de cálculo son más sencillos, además una vez calculados el Loudness específico $N'(z)$ y el Loudness total N, la obtención de S es directa como se ha descrito anteriormente. En el caso de R y F el tiempo de cálculo es mucho más elevado puesto que los algoritmos son también mucho más complejos al incorporar múltiples etapas de filtrado y no tener cálculos comunes a ambos. Una vez obtenidos los parámetros N, S, R y F, el cálculo del indicador general de molestia PA es directo, por lo que contamos con un total de 5 parámetros psicoacústicos a la salida de nuestro sistema que nos proporcionan una descripción precisa de la molestia ocasionada por los sonidos que se están emitiendo en un entorno concreto y permite realizar análisis precisos de diversos soundscapes.

Como se verá más adelante, la imposibilidad de realizar estos cálculos en un tiempo razonable en los diferentes equipos informáticos testeados ha sido uno de los motivos para buscar soluciones que mejoren este proceso en la inteligencia artificial y en concreto en las redes neuronales convolucionales.

2.1.2. Parámetros acústicos de sala

El objeto de los parámetros acústicos de sala es caracterizar el comportamiento acústico de un recinto, tal y como describíamos al comienzo del presente capítulo. Estas características acústicas son propias del espacio estudiado y dependen fundamentalmente de la geometría y los materiales que lo componen. De esta manera analizando una serie de parámetros acústicos podremos determinar cómo se va a comportar el sonido emitido en una sala y, si es necesario actuar en consecuencia para atenuar o realzar los efectos que se produzcan en el mismo. En el caso de espacios relacionados con la producción musical o la divulgación oral, los diseños arquitectónicos se realizan expresamente para el fin deseado, mediante estudios acústicos específicos, sin embargo, tanto durante el proceso de edificación, como durante el tiempo de su vida útil, el cambio de materiales constructivos, la modificación de geometrías o la inclusión de elementos externos hacen imprescindible la realización de estudios acústicos en el recinto de forma periódica para asegurar un correcto comportamiento acústico. Este comportamiento acústico se ve afectado también por la presencia humana, presentando diferentes características en función de la audiencia presente, por lo que más allá de la realización de estudios acústicos puntuales, disponer de un sistema de monitorización sencillo y rápido que permita el cálculo de parámetros acústicos de sala mientras varía la ocupación de la misma, es de gran interés y centrará parte de la investigación de esta Tesis.

Así pues, partiendo de las investigaciones de W.C. Sabine [37] sobre tiempo de reverberación o RT como parámetro acústico propiamente dicho, durante lo últimos 100 años de investigación documentada han ido tomando forma diversos parámetros acústicos de sala. En este caso no existe un modelo matemático que mediante una combinación de parámetros permita calcular la calidad acústica de una sala, tal y como sucedía con el modelo de molestia psicoacústica de Zwicker. Además debido al gran número de parámetros existentes y a su variada naturaleza, los investigadores suelen seleccionar un conjunto discreto de parámetros orientados al análisis acústico específico que se va a realizar. Este conjunto de parámetros seleccionado, se valora de forma global y permite obtener una descripción precisa del comportamiento acústico del recinto, general o enfocado a un aspecto acústico concreto, como puede ser la reverberación, la transmisión del habla o la distribución energética del sonido en la sala y que dependerá, como se ha mencionado, de sus características geométricas y constructivas.

Tal es el caso de la escuela de Gottingen [55, 56], que emplea un conjunto de tres parámetros: el tiempo de reverberación o RT (*Reverberation Time*), el índice de correlación cruzada interaural o $IACC$ (*Inter Aural Cross Correlation*) y la claridad, C. Los estudios de Yamamoto y Suzuki [57] influenciados por los anteriores, emplean C, la fuerza sonora G y evalúan

también la distribución espacial del sonido con el índice de correlación cruzada interaural ($IACC$), como en el caso anterior. En los estudios de Bradley-Soulodre [58, 59] y Barron [60] se evalúa principalmente cómo influye la energía sonora tardía en los efectos de percepción lateral del sonido mediante 2 parámetros, el nivel de energía lateral tardía GL_L (*Late Lateral Energy Level*) y la fracción de energía lateral tardía o LLF (*Late Lateral Energy Fraction*). Estos dos últimos parámetros guardan una correlación directa, pero este efecto no es extraño y se produce entre muchos de los parámetros mencionados Por lo tanto no son independientes estadísticamente hablando y una variación en uno ocasiona variaciones en otros. En relación a este punto son de especial interés los estudios de Ando [61] y Beranek [62], en el primer caso porque emplea un conjunto de parámetros estadísticamente independientes, y en el segundo porque intenta identificar qué parámetros son realmente independientes y cuáles son los más adecuados para juzgar la acústica de una sala por corresponderse mejor con la respuesta subjetiva del público. Basados en estos estudios junto con los realizados por S. Cerdá y A. Gimenez [63], los parámetros acústicos más empleados habitualmente, se pueden distinguir 4 clases según su naturaleza y la sensación subjetiva que describen:

1. Parámetros energéticos:
   - Fuerza sonora (G)
   - Claridad del habla ($C50$)
   - Claridad musical ($C80$)
   - Tiempo central (Ts)
2. Parámetros de reverberación:
   - Tiempo de reverberación (RT)
   - Tiempo de decaimiento temprano (EDT de *Early Decay Time*)
   - Brillo (Br)
3. Parámetros de inteligibilidad:
   - Índice de transmisión del habla (STI de *Speech Transmission Index*)
   - Índice simplificado o rápido de transmisión del habla, ($RASTI$ de *Rapid STI*)
   - Pérdida consonante ($\%ALcons$)
4. Parámetros de percepción espacial (llamados también espaciales):
   - Índice de correlación cruzada interaural ($IACC$ de *Inter Aural Cross- Correlation*)
   - Fracción de fuerza lateral (temprana) (LF_E de *Early Lateral Energy Fraction*)
   - Fracción de fuerza lateral coseno (temprana) (LFC_E)
   - Nivel de energía (tardía) (GL_L de *Late Lateral Energy Level*
   - Fración de energía lateral tardía (LLF de *Late Lateral Energy Fraction*)

En nuestro caso, basándonos en este amplio abanico de parámetros acústicos de sala, hemos definido un conjunto de 5 siguiendo los criterios que se exponen a continuación. La mayoría de los parámetros mencionados necesitan del cálculo previo de la respuesta impulsiva acústica de la sala, empleando al menos una fuente sonora y un micrófono con características especiales. El caso de los parámetros relacionados con la percepción espacial es singular además, pues para obtenerlos se necesita emplear un micrófono estéreo o binaural, normalmente incluidos en una cabeza de muñeco (*dummy*) que simula la humana generalmente con con orejas sintéticas y en el mejor de los casos parte de un torso, para reproducir los efectos

que producen estas partes de la anatomía humana en la percepción espacial del sonido. Debido a las dificultades que plantea este tipo de mediciones desechamos como punto de partida este tipo de parámetros, pues el fin de esta investigación es desarrollar un sistema de medición fácil de instalar y que permita obtener análisis acústicos rápidos.

Centrándonos por lo tanto en las 3 primeras clases de parámetros, los parámetros energéticos, de reverberación y de inteligibilidad, escogimos inicialmente 3 estadísticamente independientes como más representativos: el tiempo de reverberación, $RT60$, la claridad musical $C80$ y el índice de transmisión del habla STI. Para tener una mejor descripción de la distribución energética, orientada hacia al habla incluimos también la claridad del habla $C50$, que ayudó en la circunstancia que se describe a continuación. Las medidas de distancia social y reducción de aforo adoptadas en los espacios docentes debido a la pandemia del COVID-19, provocó la necesidad de realizar diversos estudios acústicos enfocados a la percepción del habla en diferentes salas de la Escuela Técnica Superior de Ingeniería de la Universidad de Valencia (ETSE-UV). En unos casos debido a que las salas habilitadas para docencia no estaban diseñadas para este fin y en otros casos debido a la distancia existente entre la zona donde se suele ubicar el profesorado y los alumnos más distantes. En este contexto la claridad del habla $C50$ ayuda a evaluar si la distribución energética del sonido en la sala favorece al habla, como se verá más delante, pero orientó nuestra atención a un parámetro acústico que no está contemplado en la lista anterior, el índice de inteligibilidad del habla o SII (*Speech Intelligibility Index*). Este parámetro nos parece muy interesante porque se obtiene directamente a partir de una señal de voz sin necesidad de calcular previamente la respuesta al impulso de la sala. El SII también refleja la dependencia de la distancia al orador y de la posición de escucha en la sala, además de revelar fenómenos de enmascaramiento, lo que lo hace muy interesante a la hora de estudiar posiciones conflictivas en un recinto.

De esta manera el conjunto de 5 parámetros acústicos de sala queda definido con: $RT60$, $C50$, $C80$, STI y SII. Este conjunto proporciona una descripción general de la acústica de la sala, pero balanceada hacia el estudio de la transmisión e inteligibilidad del habla, que ha ocupado gran parte de esta sección de la investigación. Sin embargo, como se ha descrito anteriormente, es habitual esta situación de enfocar los análisis acústicos a la faceta que más interese del análisis acústico a realizar en cada momento.

Otro aspecto a tener en cuenta a la hora de realizar análisis acústicos en salas es la normativa aplicable a los mismos, en nuestro caso relativa a los 5 parámetros del conjunto seleccionado. En el caso de la normativa técnica aplicada a los parámetros de sala más comunes, la norma ISO 3382: *Acoustics, Measurement of room acoustic parameters* [38], mencionada al principio del capítulo, es esencial a la hora de fijar el estándar de cálculo y los métodos de medida de estos parámetros acústicos, que en nuestro caso son $RT60$, $C50$ y $C80$. Esta norma se centra en parámetros acústicos relacionados con la reverberación y la distribución energética de la señal, generalmente en todo el espectro audible, pero en algunos casos como se ha visto, son necesarios parámetros específicos de una banda concreta del espectro o que evalúan aspectos muy precisos del sonido, como pueden ser los relacionados con el habla, que permite la trasmisión oral de información. En este ámbito nos interesa la norma ISO 9921: *Ergonomics - Assessment of speech communication* [64], que define parámetros como el nivel de interferencia del habla, SIL (*Speech Interference Level*) que mide la relación entre el nivel SPL_A del habla y del ruido de fondo presente en una sala o el índice de transmisión del habla, STI (*Speech Transmission Index*), que evalúa la cantidad de información de habla que se transmite con eficacia, como se verá con más detalle. De forma similar, la ANSI/ASA S3.5-1997 [65] define el índice de inteligibilidad del habla o SII (*Speech Inteligibility Index*) que cuantifica cómo de inteligible es el habla percibida y que completa por lo tanto el conjunto de normativa aplicable a nuestro conjunto de parámetros acústicos de sala.

Pese al gran número de parámetros existentes, el tiempo de reverberación RT que men-

cionábamos como base de los parámetros acústicos de sala sigue siendo el más usado habitualmente en estudios para tener una primera aproximación al comportamiento acústico de un recinto y se emplea para evaluar investigaciones en el ámbitos tan diferentes como el arquitectónico [66, 67] o el de procesado de señal, donde expertos reconocidos como M. Vorlandër emplean la reverberación si no como base, sí como una herramienta esencial para poner a prueba diversas teorías [68, 69, 70], complementado eso sí por otros parámetros acústicos calculados mediante sistemas diseñados a medida para tal fin. Esto nos lleva a los sistemas de medida existentes en el mercado, donde es difícil encontrar equipos dedicados que proporcionen más parámetros que los niveles de LA_{eq} por bandas o en casos muy específicos el tiempo de reverberación RT, ya que como se ha descrito es uno de los parámetros acústicos más empleado. Estos equipos son a veces muy raros de encontrar y a precios muy elevados. Tal es el caso de equipos como el sonómetro de *B&K 2270* o el *Audio XL2* de NTI que proporcionan niveles L_{eq} o con distintas ponderaciones como equipos habituales, en contraposición al sistema de medida de $RASTI$ (*Rapid Speech Transmission Index*) basado en emisor y receptor *B&K Type 4225 y Type 4419*.

Los parámetros psicoacústicos que describimos en el apartado anterior se pueden calcular a partir del sonido grabado en un instante de tiempo y son propios de ese sonido como se ha visto. A diferencia de estos, los parámetros acústicos de sala que nos ocupan ahora son intrínsecos a la sala y a las características geométricas y constructivas de la misma, por lo que nos dan unas directrices de cómo será afectado cualquier sonido que se produzca en la citada sala. El cálculo de la mayoría de los parámetros acústicos de sala pasa por el cálculo previo o la obtención de la respuesta impulsiva acústica de la sala. Describiéndolo de manera elemental, la respuesta al impulso de una sala se obtiene reproduciendo un sonido de muestra en la sala a analizar, grabándolo y calculando cómo ha sido afectado por las características de la misma, generalmente en varias posiciones para tener una mejor descripción y obtener si se desea una respuesta impulsiva promedio general del local.

Por esta razón los equipos de medida habituales no suelen proporcionar el cálculo de parámetros de sala, ya que son necesarias fuentes emisoras de sonido, micrófonos e interfaces de audio, además de software específico que permita calcular inicialmente la respuesta impulsiva y después los parámetros acústicos deseados. Por lo cual fuera del ámbito de la investigación y de sistemas diseñados a medida, no es usual encontrar sistemas completos de medida de parámetros acústicos de sala y menos aun que nos puedan proporcionar un conjunto diverso de parámetros o de naturaleza tan distinta como pueden ser $RT60$ o SII en un mismo análisis y para los que se emplean tanto métodos de medición como cálculos diferentes.

A continuación se describe con más detalle cada uno de los parámetros acústicos de sala que comprenden el conjunto seleccionado: Tiempo de Reverberación 60 dB ($RT60$), Claridad de la voz ($C50$), Claridad Musical ($C80$), *Speech Transmission Index* (STI) y *Speech Inteligibility Index* (SII)

Reverberation time ($RT60$) mide el tiempo que permanece un sonido presente en una sala una vez se extingue su fuente. El parámetro que definió en un principio W.C. Sabine, mide el tiempo transcurrido desde que la fuente sonora cesa su emisión hasta que la energía acústica emitida cae 60 dB y se mide en segundos. La formula general planteada por Sabine para calcular $RT60$ (Ecuación 2.8) es aplicable si se tiene un conocimiento completo tanto de la geometría de la sala a evaluar como de los materiales que la componen, pues el cálculo se basa en el volumen de la sala en m^3 (V) y en el coeficiente general de absorción de la misma, A, que se obtiene como la suma de las superficies de cada material que componga la sala multiplicado por su coeficiente de absorción acústica α.

$$RT60 = 0{,}161 \cdot \frac{V}{A} \tag{2.8}$$

A pesar de parecer algo sencillo, en la mayoría de las salas a estudiar tanto las geometrías constructivas como los materiales empleados en las mismas pueden complicar mucho su cálculo, pues conforman volúmenes difíciles de evaluar formados por superficies que pueden estar compuestas de materiales dispares con múltiples coeficientes de absorción acústica. Las geometrías no regulares, así como la presencia de pilares y otros elementos estructurales, pueden influir también en el tiempo de reverberación. Además, cualquier otro elemento auxiliar presente en la sala como el mobiliario o el equipamiento de la misma afecta al comportamiento del sonido y por lo tanto a $RT60$. En consecuencia, este cálculo nos puede dar una idea general de $RT60$ en salas regulares y bien caracterizadas pero la obtención de un valor exacto y acorde a las condiciones concretas de una sala donde desconocemos con exactitud la naturaleza de los materiales y las cotas exactas de su geometría, pasa por realizar una medición in situ y calcular $RT60$ evaluando como cae la energía acústica en la sala a lo largo del tiempo, tal y como se ve en la Figura 2.8.

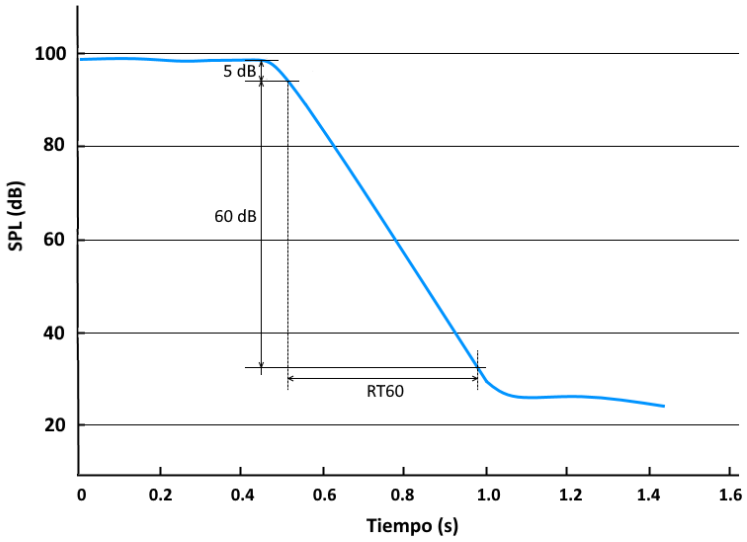


Figura 2.8: Curva de decrecimiento de energía acústica empleada para medir $RT60$.

Aunque teóricamente se busca una diferencia de 60 dB en la energía de la señal, en la práctica, la presencia de ruido de fondo puede dificultar la ubicación del nivel inferior, por lo que en algunos casos se evalúa un rango dinámico menor, empleándose $RT30$ o incluso $RT20$, que son los tiempos de reverberación con caída 30 dB y 20 dB respectivamente.

Como lo que interesa en este parámetro es la pendiente de caída, los valores de $RT30$ o $RT20$ se extrapolan para obtener $RT60$. El método de medición y cálculo de $RT60$ está definido en la norma ISO 3382-Parte 2 [39], donde se especifica que la medición comienza cuando la señal cae 5 dB por debajo de su nivel inicial o régimen estacionario, hasta caer 65 db por debajo de este, o 25 dB y 35 dB por debajo si hablamos de $RT20$ y $RT30$ respectivamente. Este método, que hemos seguido en nuestro caso, se basa en el método propuesto por Schroeder en el artículo "New method of measuring reverberation time"[71], orientado a calcular $RT60$ de manera práctica en salas donde no se podía deducir claramente una diferencia de 60 dB, a partir de la observación de respuesta impulsiva integrada temporal de la sala.

Si se observa la respuesta impulsiva de una sala, esta presenta variaciones grandes en su respuesta temporal, que la alejan del perfil ideal mostrado en la Figura 2.8, lo que obliga a realizar algún tipo de promediado que haga evidente la tendencia y obtener una curva promedio de decrecimiento de energía con la que operar. Schroeder demostró que se puede obtener la curva promedio de decrecimiento de energía integrando las contribuciones energéticas asociadas a la respuesta impulsiva de la sala, tal y como se ve en la Ecuación 2.9. Aquí se define, de forma resumida, como la curva promedio de decrecimiento de energía $\langle s^2(t) \rangle$ es igual a la integración inversa de la respuesta impulsiva $h(n)$ al cuadrado, donde N representa la potencia del ruido por banda.

$$\langle s^2(t) \rangle = N \cdot \int_t^\infty h^2(x) \mathrm{d}x \tag{2.9}$$

En el dominio discreto, la integración inversa representa la suma acumulativa de la energía y permite obtener un perfil mucho más adecuado para trabajar, como se puede ver en la parte inferior de la la Figura 2.9. En la parte izquierda de la curva integrada, una vez comienza a descender la energía, se distingue una zona donde la pendiente es más pronunciada, que nos puede proporcionar otro parámetro complementario al $RT60$ que no vamos a emplear en nuestro caso, llamado "tiempo de decaimiento temprano" o EDT (*Early Decay Time*). Una vez establecido el punto donde comienza la cola de reverberación, se puede aproximar la curva de decrecimiento de energía a una recta, empleando por ejemplo regresión lineal, como contempla la normativa.

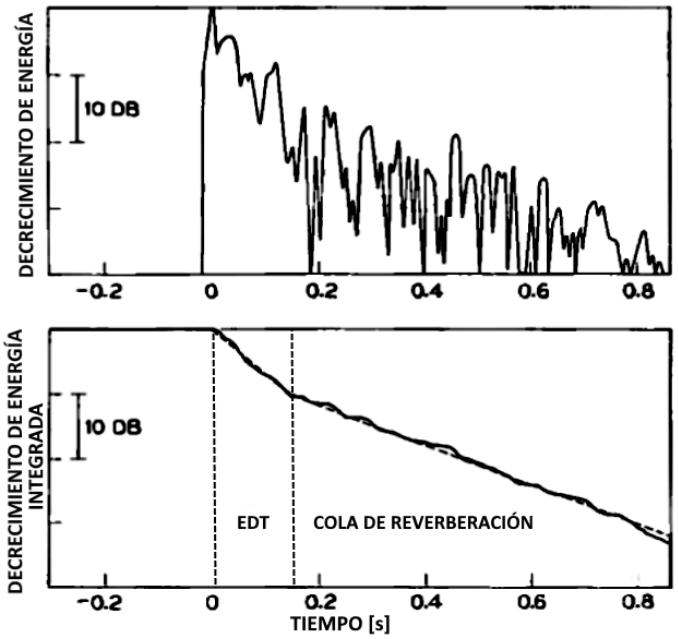


Figura 2.9: Curva de decrecimiento de energía integrada empleando el método de Schroeder.

Aunque se da libertad para elegir cualquier método que logre resultados similares, empleando el método de Schroeder que se basa en la respuesta impulsiva global integrada de la sala y siguiendo los ejemplos mostrados en la parte 2 de la ISO 3382, se puede aproximar la curva de decrecimiento de energía a una recta empleando regresión lineal, según la Ecuación 2.10a, donde a es el punto de intersección de la recta (en dB) y b la pendiente (en dB/s) con t_i siendo el tiempo en segundos de la muestra i. Aplicando un ajuste por mínimos cuadrados, se puede estimar a y b mediante las Ecuaciones 2.10b y 2.10c respectivamente, empleando los valores promedio $\bar{L}$ y $\bar{t}$, para obtener al final

$RT60$ (Ecuación 2.10f), tal y como se ha implementado en el algoritmo diseñado.

$$\hat{L}_i = a + bt_i \tag{2.10a}$$

$$a = \bar{L} - b\bar{t} \tag{2.10b}$$

$$b = \frac{\sum_{i=1}^{n}(t_i L_i) - m\bar{t}\bar{L}}{\sum_{i=1}^{n}(t_i^2) - m\bar{t}^2} \tag{2.10c}$$

$$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i \tag{2.10d}$$

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{2.10e}$$

$$RT60 = \frac{-60}{b} \tag{2.10f}$$

Cabe destacar también que en el cálculo de $RT60$ la ubicación del punto donde comienza a decaer la energía del sonido que empleamos como punto de partida para evaluar la pendiente de caída, se fija en 5 dB por debajo de su nivel estacionario pero se puede ubicar a más distancia de este en caso de que la energía no decaiga de forma muy abrupta. Por ejemplo se puede tomar un punto 10 dB por debajo del nivel estacionario, siempre que nos permita establecer un tramo lineal y medir una caída mínima de 20 dB como contempla el método de Schroeder y también la norma ISO 3382.

En la norma ISO 3382 se definen tanto los métodos de obtención de la respuesta impulsiva de la sala como los dispositivos adecuados para tal fin y las posiciones de medición, pues se debe buscar las posiciones del recinto que puedan ocasionar diferencias en los valores de $RT60$. A nivel de cálculo, tal y como hemos descrito, obtenida la respuesta impulsiva de la sala según indicaciones de la ISO 3382-1 [38] se puede extraer la curva de decrecimiento de energía por bandas de 1 octava, de 1/3 de octava o promediado general para todo el espectro, empleando el método de Schroeder, a partir de la respuesta impulsiva integrada, según se desee de definición en el cálculo. Cuando se emplean bandas de octava se promedia el valor para las bandas de 500 Hz y 1000 Hz, para el caso de bandas de 1/3 de octava se emplean las bandas que van de 400 Hz a 1250 Hz y en ambos casos el valor medio de $RT60$ devuelto se denomina T_{mid}.

En nuestro caso, para la obtención de $RT60$, se ha implementado el cálculo tanto por el método de Schroeder sobre la respuesta impulsiva global integrada, como por bandas de frecuencia, empleando las indicaciones de la ISO-3382-1 y en ambos casos buscando la aproximación a una recta de la curva de decrecimiento de energía. No obstante, por eficiencia de cálculo se ha seleccionado el primer método (que es 6 veces más rápido) y se emplea en todas las pruebas realizadas en adelante, pues interesaba tener una cifra representativa de la reverberación media de la sala, aunque reiteramos, el código empleado e incluido en el sistema de medida que se describe en la siguiente sección permite cualquier variante.

Speech Clarity ($C50$) es una medida objetiva de la claridad del sonido en una sala, en concreto relacionada con la claridad del habla. Las reflexiones tardías son desfavorables a la comprensión oral puesto que provocan la fusión de fonemas haciendo que el habla sea poco clara. No obstante, si el retraso no supera un determinado límite de tiempo estas contribuirán de manera positiva a la comprensión del habla. Este límite de tiempo que separa las reflexiones beneficiosas de las perjudiciales para el caso del habla es de 50 ms. De esta manera, $C50$ cuantifica cuanta energía de la respuesta impulsiva está

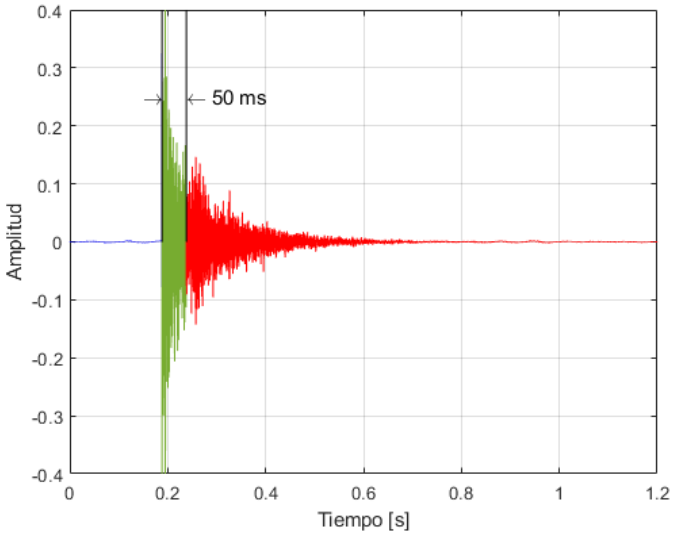


Figura 2.10: Respuesta impulsiva de una sala. Intervalos empleados en el cálculo de $C50$.

en el tramo de interés (los primeros 50 ms) frente a la que está fuera de este intervalo, como se puede ver gráficamente en la Figura 2.10, en verde y rojo respectivamente.

$C50$ está recogido en la ISO 3382-1 y se calcula mediante la Ecuación 2.11 como el cociente entre la energía sonora registrada en los primeros 50 ms desde la llegada del sonido directo y la que llega después de los 50 ms, expresándose el resultado en dB.

$$C50 = 10 \cdot log \frac{\int_0^{50ms} p^2(t)\mathrm{d}t}{\int_{50ms}^{\infty} p^2(t)\mathrm{d}t} \tag{2.11}$$

El empleo de una respuesta impulsiva promediada por bandas nos proporciona un valor de $C50$ general pero usualmente se realizan análisis por bandas de 1 octava entre 125 Hz y 4 kHz. Para obtener un valor más representativo del $C50$ de una sala se suele emplear el *C50 Speech Average* ($C50_{SA}$) propuesto por L.G. Marshall [72], calculado a partir de los valores de $C50$ de las bandas de 500 Hz a 4 kHz, mediante los valores de ponderación empleados en la Ecuación 2.12.

$$C50_{SA} = 0{,}15 \cdot C50_{500Hz} + 0{,}25 \cdot C50_{1kHz} + 0{,}35 \cdot C50_{2kHz} + 0{,}25 \cdot C50_{4kHz} \tag{2.12}$$

Como norma general para tener una claridad aceptable del sonido relacionado con el habla se necesita un $C50$ mayor o igual a 2 dB.

Musical Clarity ($C80$) nos proporciona una medida objetiva de la claridad o grado de separación con la que percibimos los sonidos, en este caso relacionados con la música. Como en el caso de $C50$, las reflexiones tardías que superan cierto límite de tiempo contribuyen negativamente a la percepción nítida de la música. En el caso de $C80$, son los primeros 80 ms, ya que el oído integra las reflexiones recibidas en ese intervalo como pertenecientes al sonido directo, reforzando el mismo, pero las percibidas después de este umbral contribuyen a distorsionarlo y a disminuir la claridad musical. Así, como se puede ver en la Figura 2.11 sobre una respuesta impulsiva en este caso como energía en función del tiempo, $C80$ cuantifica la proporción de energía de la respuesta impulsiva presente en los primeros 80 ms desde la llegada del sonido directo (representado en verde) respecto a la que está fuera de este intervalo (representado en rojo).

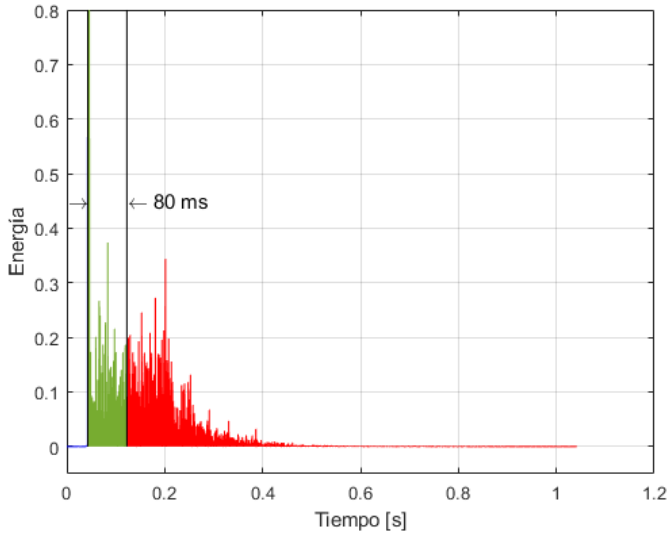


Figura 2.11: Respuesta impulsiva de una sala como energía en función del tiempo. Intervalos para calcular $C80$.

Como en el caso anterior, el cálculo de $C80$ está recogido en la ISO 3382-1, obteniéndose mediante la Ecuación 2.13, como el cociente entre la energía sonora registrada en los primeros 80 ms desde la llegada del sonido directo y la que llega después de los 80 ms, expresado en dB.

$$C80 = 10 \cdot log \frac{\int_0^{80ms} p^2(t)\mathrm{d}t}{\int_{80ms}^{\infty} p^2(t)\mathrm{d}t} \tag{2.13}$$

Aunque se puede calcular $C80$ a partir de una respuesta impulsiva promedio, no representa fielmente el comportamiento de la sala por lo que se suele calcular por bandas de frecuencia de 1 octava. Para obtener un único valor representativo, se emplea una media de los valores de C80 en las bandas de 500 Hz, 1 y 2 kHz (Ecuación 2.14), que se denomina *C80 Music Average*.

$$C80_{MA} = \frac{C80_{500Hz} + C80_{1kHz} + C80_{2kHz}}{3} \tag{2.14}$$

Como norma general según la ISO 3382-1, $C80$ debe estar en un rango entre -5 dB y +5 dB, aunque este intervalo es más restrictivo dependiendo del uso de la sala a estudiar. Así, para una sala dedicada a música de cámara, el intervalo ideal de $C80$ debería ir de -2 dB a 2 dB, pero para una dedicada a la ópera debería estar al menos entre 1 dB y 3 dB lo que significa que necesitaremos más energía de la señal reforzando el sonido directo para percibir más claramente la ópera.

Speech Transmission Index (STI) mide la calidad de la transmisión oral de la información. Es un parámetro acústico complejo que analiza la respuesta del canal acústico entre el hablante y el oyente, ya sea una sala, un equipo electrónico o una línea telefónica. En función de la respuesta de este canal se estima cuánta información podrá transmitirse de manera eficaz. Hay diversos factores que pueden afectar al STI, las características de la sala como el tiempo de reverberación $RT60$ o los ecos (reflexiones tardías), el nivel de ruido de fondo, el volumen de la voz del hablante o efectos psicoacústicos de enmascaramiento si la transmisión es oral. Si el hablante se ayuda de un medio de amplificación o el canal es un medio electrónico se pueden sumar además efectos producidos por la calidad del equipo de reproducción y distorsiones no lineales

en el mismo. El concepto del STI fue desarrollado por T. Houtgast y H. Steeneken en 1971 y aprobado por la *Acoustical Society of America* en 1980 [73], cuando trabajando para las Fuerzas Armadas de los Países Bajos, establecieron un método objetivo para evaluar como variaba la inteligibilidad del habla dependiendo del canal de comunicación empleado, ya fuera un sistema de radio o una sala de reuniones. Así pues, STI está basado en cuantificar cómo se deterioran las fluctuaciones o modulaciones de la intensidad del habla debido al canal de comunicaciones, como se puede ver en la Figura 2.12. Caracterizando la variación de estas fluctuaciones en función del tiempo por bandas de frecuencia se obtienen las funciones de transferencia de modulación o *Modulation Transfer Function*, en adelante MTF, cuyo análisis sirve para calcular STI. Originalmente, empleando señales artificiales, se medía cómo variaba la modulación de las $MTF'S$ mediante 14 señales moduladoras aplicadas a cada banda de octava entre 125 Hz y 8 kHz (7 bandas $\times$ 14 señales moduladoras = 98 señales), comparando entre las señales emitidas y las recibidas. Se obtenían a continuación varias señales-ruido aparente, que se ponderaban para obtener el valor de STI.

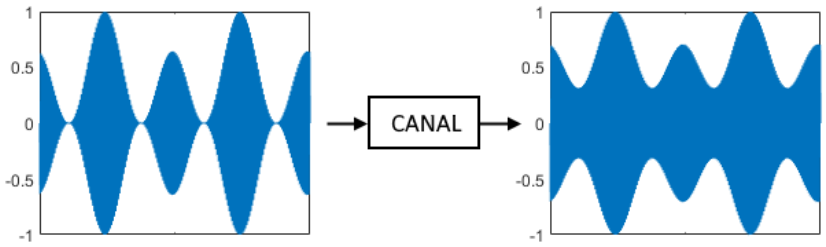


Figura 2.12: Cambio de profundidad de modulación producida por el canal de transmisión.

Debido a la complejidad del cálculo apareció una simplificación para ser aplicada únicamente a salas, llamada $RASTI$ de *Room Acoustics STI*, cuyo acrónimo interpretan también algunos autores como *Rapid Speech Transmission Index*. Para obtener el $RASTI$ sólo se analizan las bandas de 500 Hz con 4 frecuencias moduladoras y de 2 kHz con 5 señales moduladoras lo que resulta únicamente en 9 señales en total. Se hizo popular empleándose en muchos análisis hasta desarrollarse el $STIPA$ de *STI for Public Address systems* como aproximación a STI, que lo sustituyó en el año 2000 debido a su velocidad de cálculo similar y su mayor precisión, que se acerca a la del STI completo, siendo declarado obsoleto $RASTI$ en 2011 por la IEC (*International Electrotechnical Commission*). El método de medida, así como las señales y cálculos necesarios para obtener STI están definidos en la norma IEC-60268-16 [[74]], que hemos seguido en nuestro caso.

STI se calcula de forma indirecta a partir de la respuesta impulsiva de la sala filtrada por bandas de octava, desde 125 Hz a 8 kHz (7 bandas). Se denomina forma indirecta porque las funciones de transferencia de modulación $MTF'S$ se calculan a partir de la respuesta impulsiva y no se miden empleando diferentes frecuencias moduladoras. Esto es posible siempre que se aplique a respuestas impulsivas de sistemas lineales, pasivos e invariantes en el tiempo. Se consigue así hacer más robusto el cálculo frente a efectos introducidos por el sistema de reproducción y grabación de las señales moduladas.

De esta manera, lo primero que se hace es calcular a partir de la respuesta impulsiva filtrada por las 7 bandas de interés, las MTF de cada banda, $m(F_0)$ (Ecuación 2.15a). En segundo lugar se modula cada $m(F_0)$ con cada una de las 14 frecuencias moduladoras F_m de 0.63 a 12.5 Hz, que cubren las modulaciones habituales del habla, y se calcula la reducción del índice de modulación por banda y frecuencia moduladora $m(F_0, F_m)$

(Ecuación 2.15b).

$$m(F_0) = \frac{\int_0^\infty h_f^2(\tau) \cdot exp(-j \cdot 2\pi \cdot F \cdot \tau)\mathrm{d}\tau}{\int_0^\infty h_f^2(\tau)\mathrm{d}\tau} \tag{2.15a}$$

$$m(F_0, F_m) = m(F_0) \otimes F_m \tag{2.15b}$$

A continuación se obtiene la relación señal-ruido aparente de cada índice (F_0, F_m), denotada por $SNR_{ap}(F_0, F_m)$ mediante la Ecuación 2.16a. Los 98 valores obtenidos son truncados para que estén dentro del intervalo [-15, 15] dB de forma que el valor final de STI esté en el intervalo [0, 1]. Posteriormente, se promedian por banda para obtener $\overline{SNR}_{ap}(F_0)$ (Ecuación 2.16b), que es la relación señal-ruido aparente por bandas F_0.

Finalmente, se ponderan estos valores mediante la función $g(F_0)$ para obtener una relación señal-ruido aparente general mostrada en la Ecuación 2.16c, que se emplea para calcular el valor final de STI según la Ecuación 2.16d.

$$SNR_{ap}(F_0, F_m) = 10 \cdot log(\frac{m(F_0, F_m)}{1 - m(F_0, F_m)}) \quad \in [-15, 15]dB \tag{2.16a}$$

$$\overline{SNR}_{ap}(F_0) = \frac{\sum_{Fm} SNR_{ap}(F_0, F_m)}{14} \tag{2.16b}$$

$$\overline{SNR}_{ap} = \sum_{F_0} \overline{SNR}_{ap}(F_0) \cdot g(F_0) \tag{2.16c}$$

$$STI = \frac{\overline{SNR}_{ap} + 15}{30} \tag{2.16d}$$

El valor de STI obtenido estará comprendido entre 0 y 1, representando respectivamente de una muy mala a una excelente transmisión de la información del habla. Normalmente se divide el rango completo de valores de STI en 5 intervalos contemplados en la norma IEC-60268-16 [[74]], como se puede ver en la Tabla 2.1.

Tabla 2.1: Intervalos de STI según IEC-60268-16.

Intervalo de STI	**Calidad de transmisión**
0.00 – 0.30	Mala
0.30 – 0.45	Pobre
0.45 – 0.60	Regular
0.60 – 0.75	Buena
0.75 – 1.00	Excelente

La mínima diferencia de STI que puede detectar el oído humano se encuentra entre 0.03 y 0.04 [75], por lo que en la documentación de la norma IEC-60268-16 se puede encontrar también otra escala de 12 intervalos con diferencias de 0.04 entre cada uno y que comprende valores entre 0.36 y 0.76 de STI a los que asigna letras de A+ a U para calificar salas según el valor de STI promedio. Se emplea ese rango limitado de valores debido a que son los más usuales para salas generales de uso común donde encontrar valores de STI superiores a 0.75 no es habitual. De hecho en espacios públicos con un ruido normal y afluencia media de gente, como puede ser una estación de ferrocarril

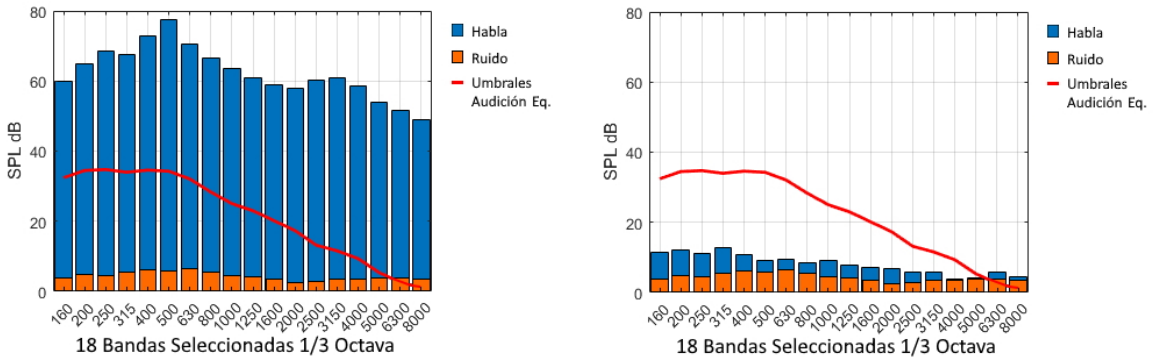
de tamaño medio, aunque se disponga de un sistema de megafonía, los valores de STI suelen estar entre 0.45 y 0.60.

Como STI evalúa las características físicas del canal de comunicaciones y su capacidad de transmitir señales que contengan patrones característicos del habla, es independiente del idioma empleado y nos dará un indicador fiel de la calidad de la transmisión de la información del habla que permite un canal, sala o sistema de comunicaciones. Si se desea evaluar además cómo de inteligible es una señal de voz en concreto, contemplando efectos como ruido presente o características perceptivas del oyente, se emplea el parámetro Speech Intelligibility Index o SII, que complementa al STI y hemos incluido en el conjunto de parámetros de sala a emplear, como se describe a continuación.

Speech Intelligibility Index (SII) mide la proporción de la información del habla que es audible y utilizable por el oyente. Hasta hace poco tiempo era más conocido otro parámetroo, el Índice de Articulación o en inglés, *Articulation Index*, AI en adelante. El AI fué desarrollado desde 1947 y estandarizado en 1969 en la norma ANSI/ASA S3.5. A partir de entonces se fueron introduciendo mejoras al cálculo de AI hasta que se llegó al SII, que sustituyó al anterior en la revisión del estándar de 1997, pasando a llamarse *ANSI-S3.5:1997, Methods for Calculation of the Speech Intelligibility Index* [65]. Aunque los dos métodos se basan en la teoría de que la inteligibilidad del habla está directamente relacionada con la proporción de información audible del habla, existen ciertas diferencias importantes entre el antiguo AI y el SII. Estas se basan en que el SII habilita un marco más general para realizar los cálculos, permitiendo flexibilidad en las variables de entrada, como pueden ser la inclusión de niveles del habla, la definición de umbrales de ruido y umbrales de audición. De manera paralela al STI, los valores de SII se mueven entre los valores 0 cuando ninguna información del habla es entendible, hasta 1 cuando toda la información del habla es audible y utilizable para un oyente en un entorno determinado. Sin embargo SII no se calcula a partir de la respuesta impulsiva de la sala, sino que se hace a partir de la señal del habla filtrada por bandas a la que se pueden incorporar medidas de ruido si está presente en la sala además de curvas de sensibilidad acústica del oyente. El hecho de que el algoritmo contemple tanto la inclusión de medidas de ruido como de patrones de escucha fuera de las curvas características de percepción acústica del oído, ha permitido emplear el SII para evaluar entornos donde la comunicación es vital pese a la presencia de ruido, como puede ser un quirófano o entornos donde se emplea protección auditiva que puede modificar el patrón de escucha pero permitir la comunicación, como entornos industriales.

De un modo esquemático, SII se calcula determinando qué proporción de una señal de habla es audible en determinadas bandas de frecuencia. Siguiendo las indicaciones del estándar ANSI/ASA-S3.5, como se ha hecho en nuestro caso, para calcular el SII necesitaremos 4 elementos: información sobre el espectro de la señal de habla evaluada, información espectral de los niveles de ruido (si está presente), el umbral auditivo a emplear e información sobre la importancia por bandas del habla evaluada.

En la Figura 2.13 se pueden ver dos ejemplos muy extremos de niveles SPL empleados para el cálculo de SII. En estos ejemplos, se han considerado una división por bandas de 1/3 de octava para tener una buena definición. En ambos casos los niveles de ruido (barras naranja) y los niveles equivalentes de audición (línea roja) son los mismos, sin embargo los niveles SPL de la señal de voz varían sustancialmente de un caso a otro, lo que dará a lugar a un SII elevado en el caso de la Figura 2.13a y de un SII muy bajo en el caso de la Figura 2.13b. De esta manera se ve claramente cómo el hecho de tener una señal de voz pequeña en proporción al ruido ambiente presente o que este nivel de la señal de voz esté muy por debajo de los niveles equivalentes de audición contemplados



(a) Señal de habla con SII alto

(b) Señal de habla con SII bajo

Figura 2.13: Niveles SPL de habla, ruido y niveles equivalentes de audición para el cálculo de SII.

en la norma, provocan un valor de SII bajo, lo que se puede traducir en una muy baja inteligibilidad del habla. Además de las tres magnitudes mencionadas: el nivel de habla, nivel de ruido y nivel equivalente de audición, en el calculo del SII interviene un cuarto elemento que pondera los demás, que es la función de importancia de banda o *Band importance function*, mostrada en la Figura 2.14 con los valores promedios para el idioma inglés según la norma ANSI/ASA-S3.5, pues al estar basadas en estímulos de habla específicos, se definen varias curvas para diferentes idiomas, edad o sexo del hablante, y la norma permite emplear la más adecuada a cada situación.

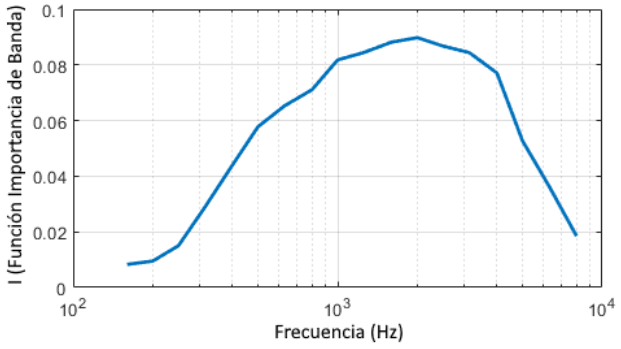


Figura 2.14: *Average Band importance function* para idioma ingles empleada en el cálculo de SII.

La expresión general para obtener SII se muestra en la Ecuación 2.17 y se basa en obtener los valores de audibilidad por banda A_i y multiplicarlos por la descrita *Band importance function*, I_i.

$$SII = \sum_{i=1}^{n} I_i A_i \tag{2.17}$$

En la Figura 2.15 se detalla el algoritmo empleado para calcular el *Speech Intelligibility Index*. El procedimiento completo de cálculo se divide en varios pasos descritos en la norma, siendo el primero de ellos la determinación del número de bandas en frecuencia a emplear. En cuanto a las bandas empleadas para el cálculo (n), la norma actual es flexible y se puede escoger el grado de detalle, desde 6 bandas si se analiza por octava, hasta 21 bandas si se hace por bandas críticas. Cuanto más bandas se contemplen, más preciso será el cálculo. En nuestro caso hemos escogido realizar un análisis por bandas de 1/3 de octava (18 bandas).

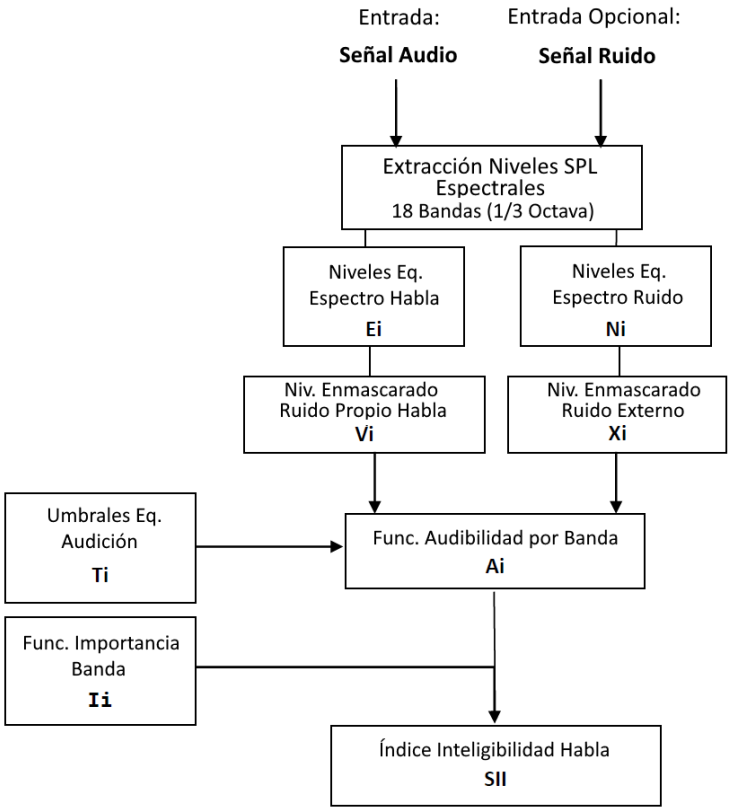


Figura 2.15: Algoritmo para el cálculo del SII.

El siguiente paso consiste en especificar 3 parámetros que se utilizarán en el cálculo: el espectro equivalente del habla E_i que se extrae de la señal de audio de entrada, el espectro equivalente del ruido N_i, extraído de la señal de ruido y los umbrales de audición T_i. A continuación se se realizan una serie de comparaciones que determinarán el enmascarado real de la señal de voz que se produce en cada banda debido al ruido propio (V_i) y externo (X_i) contrastándolos con los umbrales de audición T_i dando lugar a los valores de audibilidad por banda A_i. Para finalizar se realiza la suma del producto de A_i e I_i (función de importancia de banda), que dará lugar al SII.

El valor final de SII está en el intervalo [0, 1] y no posee unidades ya que representa la proporción de la señal del habla que es audible y utilizable por el oyente. La escala completa se puede dividir en 5 intervalos como en el caso de STI mostrado en la Tabla 2.1, pero con la diferencia de que un valor de 0.5 de SII representa que el 50 % de la información oral es audible, sin embargo en la mayoría de situaciones entenderíamos más del 80 % de las oraciones con alguna dificultad, eso sí, por lo que sería más que suficiente en una situación normal. Sin embargo estos valores de SII pueden generar un sobre-esfuerzo a la hora de tener que comprender la información oral, por lo que no son deseables en espacios donde sea vital que la información se transmita adecuadamente. Así pues, determinados espacios como puede ser un quirófano o una sala de control aeroportuaria necesitarán de unos valores de SII elevados para garantizar la inteligibilidad del habla sin esfuerzos. De la misma manera, si se garantiza esta inteligibilidad elevada en los espacios docentes mediante valores de SII elevados, se facilita en gran medida la adecuada transmisión de la información de forma oral.

En la Figura 2.16 se muestra esquemáticamente el algoritmo diseñado para la obtención del conjunto de parámetros de sala seleccionado. La calibración se ha llevado a cabo siguiendo

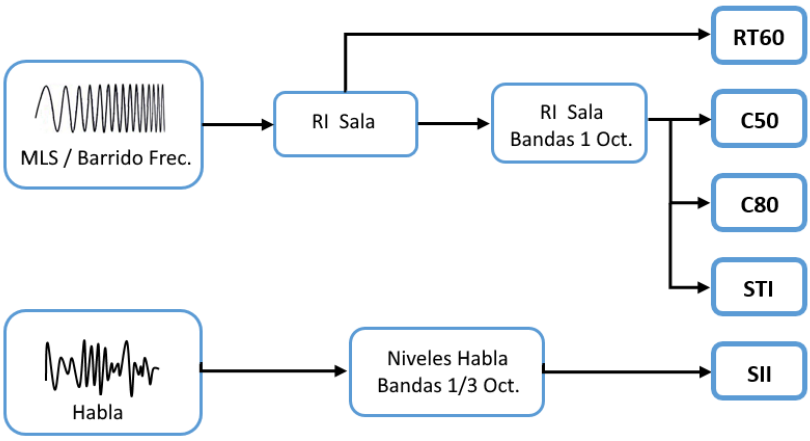


Figura 2.16: Algoritmo para la obtención del conjunto de parámetros de sala.

instrucciones de la normativa pertinente y se ha verificado mediante el software *ACQUA*,[2] *Odeon*[3] y *Matlab Audio Toolbox*[4] .

Tal y como se ha descrito a lo largo de esta sección, para obtener 4 de los 5 parámetros es necesario extraer la respuesta impulsiva de la sala. De esta manera, siguiendo indicaciones de la norma ISO-3382 se calcula primero la respuesta impulsiva mediante la emisión de secuencias de longitud máxima o MLS (*Maximum Length Sequence*). Esta respuesta impulsiva se integra para calcular $RT60$ mediante el método de Schroeder y se filtra por bandas de 1 octava para calcular $C50$, $C80$ y STI. De un modo similar pero reproduciendo una señal de habla en inglés, se procede a filtrar por bandas de 1/3 de octava la señal recibida y proceder al cálculo de SII. Los filtros utilizados para bandas de octava y 1/3 de octava cumplen con las especificaciones del estándar IEC-61260-1 [48] , siguiendo también indicaciones de la norma ISO-3382.
Se puede ver fácilmente que el proceso es mucho más complejo que con el bloque anterior de parámetros relacionados con el modelo de molestia psicoacústica de Zwicker, ya que implica la reproducción de diferentes señales y del cálculo previo de respuestas impulsivas tanto integradas, como por bandas de frecuencia, todo esto antes de proceder al cálculo de los parámetros de sala descritos.

2.2. Sistemas de medida

A la hora de evaluar uno o más parámetros acústicos es esencial disponer de un sistema de medida que permita su cálculo directo o como mínimo el almacenamiento de señales muestreadas para su posterior análisis y cálculo. En el caso de los parámetros psicoacústicos, los sonidos se producen en el momento y disponer de los valores de los parámetros en el menor tiempo posible, es muy importante para correlar eventos con niveles de molestia psicoacústica. En el caso de los parámetros acústicos de sala, los análisis comprenden un proceso de toma de medida y cálculo más extenso y complejo, por lo que disponer de un sistema que permita realizar el proceso completo y proporcione los parámetros acústicos de sala en el menor tiempo posible, es fundamental para realizar análisis acústicos de calidad. En la base de esta Tesis Doctoral está el desarrollo de nuevas herramientas de monitorización acústica más rápidas y completas que las existentes y para ello hemos comenzado analizando sistemas de medida

[2]https://www.head-acoustics.com/products/analysis-software/acqua
[3]https://odeon.dk/
[4]https://es.mathworks.com/products/audio.html

clásicos o cableados hasta llegar a sistemas distribuidos con sensores inalámbricos basados en IoT. En esta sección se describen los sistemas de medida empleados así como su diseño, desarrollo e implementación en cada caso.

2.2.1. Sistema de medida cableado

En lo que se refiere a sistemas de medida comerciales, tal y como se ha ido describiendo en la sección anterior, es difícil si no imposible encontrar uno que integre el cálculo de diversos parámetros acústicos al mismo tiempo, o al menos todos los parámetros recogidos en los dos conjuntos que se han seleccionado para esta investigación. Así pues, la mayoría de sonómetros proporcionan únicamente los niveles SPL y L_eq, como mucho con distintos valores de ponderación por bandas (A, B, C, etc), como pueden ser los sonómetros *B&K 2270-S*[5] o el *Audio XL2* de NTI.[6] Eso en el caso de parámetros relacionados con sonidos que se están produciendo en el mismo momento y que no precisan de una fuente de sonido y sólo para un conjunto muy reducido de parámetros, pues para análisis que precisan del cálculo previo de la respuesta impulsiva de una sala, la cosa se complica. En este sentido, los sistemas más complejos que se encuentran en el mercado se basan en una unidad de adquisición de datos DAQ que realiza las labores de procesado también y a la que se conecta una fuente de sonido y uno o varios micrófonos. Algunos ejemplos de la evolución de estos sistemas se muestran representados en la Figura 2.17, como el sistema de medida de $RASTI$ (*Rapid Speech Transmission Index*) basado en emisor y receptor *B&K Type 4225 y Type 4419*[7] (Fig.2.17 izquierda), el sistema de NTI Flexus FX100[8] (Fig.2.17 centro), o los sistemas a medida que propone B&K y que se basan en varios equipos individuales: una fuente de sonido (estándar u omnidireccional), micrófono de medida y un DAQ que se conecta a un ordenador personal externo donde un software de la marca se encarga de los cálculos pertinentes (Fig.2.17 derecha). Tanto el software como el hardware de los sistemas mostrados es propietario y ceñido a

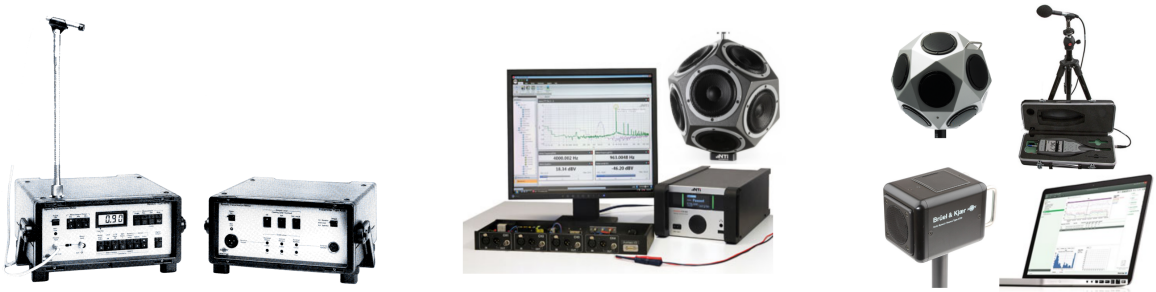


Figura 2.17: Ejemplos de sistemas de análisis acústico del mercado.

análisis muy concretos y un conjunto limitado (si es más de uno) de parámetros acústicos. El hardware y el softaware tanto si se adquieren por separado como si se hace de manera conjunta, suelen tener precios muy elevados debido a la alta especialización de los sistemas tratados, lo que complica aun más el hecho de encontrar un sistema de medida que proporcione varios parámetros acústicos. En muchos casos, para obtener cálculos de parámetros acústicos fuera del conjunto incluido por el sistema hay que hacer uso de software de terceros para el análisis de las señales adquiridas, en caso de que estas sean exportables. El precio, la dificultad de adquirir sistemas completos genéricos que permitan diversos tipos de análisis y en especial

[5] https://www.bksv.com/es/instruments/handheld/sound-level-meter-kits/sound-intensity-2270-s

[6] https://www.nti-audio.com/es/productos/sonometro-xl2

[7] https://www.bksv.com/media/doc/bo0244.pdf

[8] https://www.nti-audio.com/es/productos/flexus-fx100

la opacidad del software propietario empleado para los cálculos, ha propiciado que para las medidas realizadas en esta Tesis Doctoral se haya implementado un sistema propio basado en dispositivos comerciales y software propio verificado y calibrado según la normativa vigente.

De esta manera, se ha diseñado un sistema de análisis acústico cableado, basado en un sistema de adquisición de datos al que se conectan tanto micrófonos como fuentes de sonido, permitiendo transmitir las señales acústicas adquiridas a un ordenador personal (PC en adelante) donde se realizan los cálculos de los parámetros acústicos deseados.

- **Sistema de adquisición de datos**: Hemos empleado la interfaz de audio profesional Presonus AudioVox 1818-VSL[9] , que permite grabar y reproducir hasta 8 entradas de micrófono y 8 salidas analógicas balanceadas a frecuencias de muestreo de hasta 96 kHz, con un ruido ultra bajo, y que incorpora una interfaz USB que permite la conexión a diferentes equipos informáticos.

- **Micrófonos:** Para la adquisición de las señales de audio hemos empleado micrófonos de medición Behringer ECM8000.[10] Estos micrófonos de condensador poseen un patrón de recepción omnidireccional con una respuesta en frecuencia ultra-lineal con muy bajo nivel de ruido a la salida. Dependiendo del análisis a realizar se han conectado uno o más micrófonos de forma simultánea, aprovechando las 8 entradas de micrófono disponibles.

- **Fuentes de Sonido:** Si la medida lo requiere, en nuestro caso hemos empleado 3 fuentes con distinta directividad:

  1. Fuente omnidireccional en forma de dodecaedro con altavoces Visaton B-100[11] que hemos diseñado para este sistema expresamente.
  2. Monitor activo de estudio ESI nEar-05[12] de 5 pulgadas y 30W.
  3. Monitor multimedia activo M-Audio AV-30[13] de 3 pulgadas y 10W.

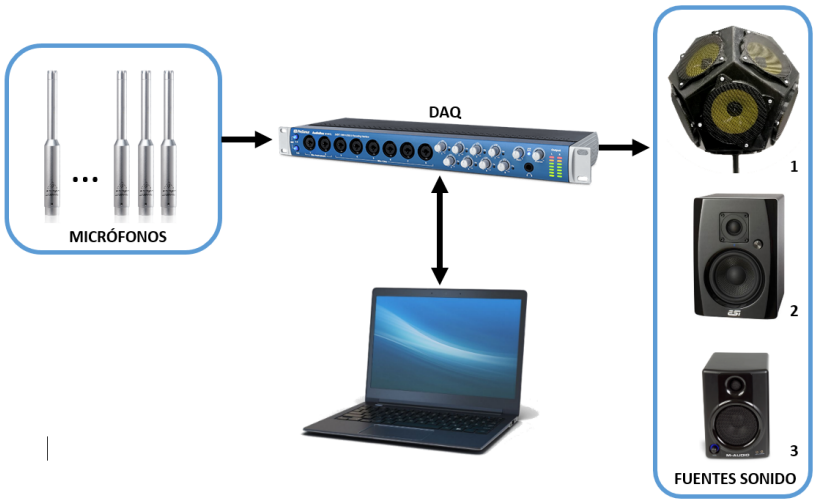


Figura 2.18: Sistema de medida cableado.

En la Figura 2.18 se muestra un esquema del sistema cableado diseñado. Dependiendo del conjunto de parámetros que se vayan a analizar se emplearán unos elementos u otros. Así,

[9]https://www.presonus.com/productos/es/AudioBox-1818VSL
[10]https://www.behringer.com/product.html?modelCode=P0118
[11]https://www.visaton.de/en/products/drivers/fullrange-systems/b-100-6-ohm
[12]https://www.esi-audio.com/products/near05classic
[13]https://m-audio.com/products/view/studiophile-av-30

si se van a monitorizar los parámetros de molestia psicoacústica, que comprenden el primer conjunto de parámetros seleccionados para nuestra investigación, se hará uso de los micrófonos ubicados en diversas posiciones de interés a monitorizar y las señales procedentes de la interfaz de audio se transmitirán a un PC donde se realizan los cálculos de los parámetros, siguiendo el esquema mostrado en la sección 2.1.1 en la Figura 2.7.

Por el contrario, si estamos en un entorno en el que se desee calcular los parámetros acústicos de sala, haremos uso tanto de los micrófonos como de las fuentes de sonido. Siguiendo el algoritmo descrito en la Sección 2.1.2 en la Figura 2.16, el primer paso pasa por calcular la respuesta impulsiva de la sala empleando la fuente omnidireccional o el monitor ESI nEar05 si se desea algo más de directividad en la fuente. Los barridos en frecuencia o las señales MLS reproducidas son grabadas por un micrófono o varios si deseamos calcular respuestas impulsivas en distintos puntos o bien promediar una y transmitidas al PC para realizar el cálculo de la respuesta impulsiva de la sala y posteriormente de los parámetros $RT60$, $C50$, $C80$ y STI. Para calcular el parámetros SII se usa una fuente direccional como el monitor M-Audio AV-30 para simular un orador en la sala y recoger las señales de habla reproducidas mediante la ubicación de los micrófonos en las posiciones que deseemos monitorizar. Las señales grabadas mediante los micrófonos, como en los casos anteriores se transmiten al PC donde se realiza el cálculo de SII para cada posición. En la Figura 2.19 se pueden ver algunos ejemplos empleando el sistema cableado en distintos escenarios para calcular diferentes parámetros acústicos. La calibración del sistema a nivel hardware se ha llevado a cabo mediante un calibrador CESVA CB-5 mediante señal de 1 KHz a 94 y 104 dB y mediante un sonómetro CESVA SC310. A nivel de software, como ya se ha comentado en cada caso, se han empleado los procesos de calibración descritos en la norma pertinente de cada caso y verificado empleando los programas informáticos *ArtemiS Suite*[14] , *ACQUA*[15] , *Odeon*[16] y *Matlab Audio Toolbox*[17] .

Tanto para el caso de los parámetros de molestia psicoacústica como para el caso de los parámetros acústicos de sala, si se desea estudiar un área de dimensiones considerables, se deduce rápidamente que la instalación o montaje del sistema de medida cableado no es sencillo. Para tener un muestreo acústico aceptable del entorno será necesario ubicar suficientes micrófonos y las distancias entre ellos, la interfaz de audio y las fuentes de sonido, en caso de ser necesaria, pueden dificultar la instalación del sistema si no hacerla imposible por la longitud de cables necesaria. En muchos casos será necesario realizar la medida en diferentes etapas trasladando el sistema completo lo que ralentiza y dificulta en gran medida el estudio acústico de los entornos. En el caso de un soundscape de una zona urbana, como puede ser un barrio completo, el empleo de un sistema cableado es generalmente inviable si no se ha previsto en la urbanización inicial.

2.2.2. Red de sensores inalámbricos y sistema IoT de monitorización

Como hemos visto, el despliegue de un sistema de medida acústica cableado en una zona de dimensiones considerables puede plantear dificultades. Por lo tanto, a la hora de desarrollar nuevos sistemas de monitorización acústica es lógico enfocar nuestra atención en los sensores inalámbricos como solución a este problema. El uso de sensores inalámbricos facilita no solo la instalación sino el incremento de puntos de muestreo y la variación rápida de posiciones de los mismos. Además, gracias al reducido consumo de los nodos empleados, se puede hacer uso de baterías eliminando la necesidad de estar conectado a la red eléctrica y convirtiendo

[14]https://www.head-acoustics.com/products/analysis-software/artemis-suite

[15]https://www.head-acoustics.com/products/analysis-software/acqua

[16]https://odeon.dk/

[17]https://es.mathworks.com/products/audio.html

Figura 2.19: Ejemplos de uso del sistema de medida cableado.

a los sensores en inalámbricos completamente. Estas características nos han conducido a desarrollar el sistema de monitorización acústica basado en una red de sensores acústicos inalámbricos que se describe en esta sección.

De hecho, las redes de sensores acústicos inalámbricos, o *Wireless Acoustic Sensor Networks*, en adelante WASN, han recibido en los últimos años una atención creciente en diferentes campos de investigación, como en el procesado de señal acústica, la interacción hombre-máquina, la investigación sobre aprendizaje automático, etc. Ya en el campo de la acústica, gracias a las WASN se ha dado enfoques novedosos a la resolución de problemas clásicos, como la localización de fuentes acústicas [8, 9], la detección y clasificación de eventos acústicos o la evaluación del ruido ambiental [11]. Otros campos como la asistencia doméstica han sido revisados también asumiendo diferentes escenarios y considerando nodos acústicos inalámbricos [12]. La monitorización de soundscapes tanto en el ámbito de la contaminación acústica como en el desarrollo de redes de sensores económicos o low-cost, ha sido susceptible también de adoptar tecnologías basadas en WASN mediante nodos conectados a internet o nodos IoT, como puede verse en los artículos [76, 77, 78].

Así pues, para resolver los problemas de logística y tiempo que plantea la instalación de

un sistema de medida cableado, diseñamos un sistema basado en una WASN que permita la monitorización de diferentes parámetros acústicos de forma inalámbrica, tanto de forma independiente como conectado a internet empleando tecnología y protocolos IoT. En la Figura 2.20 se muestra un esquema del sistema diseñado.

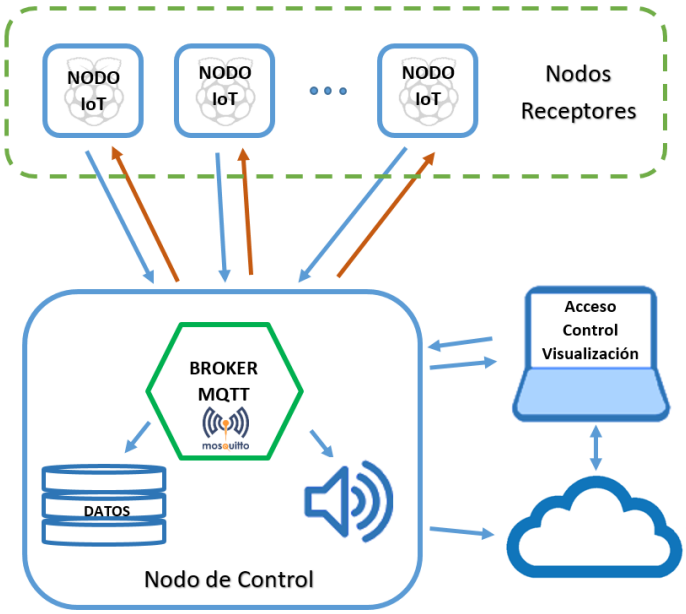


Figura 2.20: Sistema de medida inalámbrico basado en IoT.

El principio que se ha seguido para diseñar este sistema ha sido conseguir la máxima versatilidad posible. Está basado en varios nodos receptores que adquieren el audio mediante micrófonos y de un nodo de control que realiza diferentes funciones: punto de acceso inalámbrico, fuente de sonido, base de datos, agente de control y centro de cálculo en el caso de que los parámetros acústicos no se calculen en los nodos, como se verá más adelante.

Durante la etapa de pruebas y desarrollo se han evaluado numerosas configuraciones y la que ha resultado presentar más ventajas además de ser la más robusta ha sido con el nodo de control actuando como punto de acceso inalámbrico, base de datos y agente de control, volcando los parámetros calculados a internet si es posible la conexión, cosa que no es posible en muchos escenarios. El hecho de implementar la función de punto de acceso inalámbrico en el nodo de control permite además menos latencia a la hora de enviar datos y una mejor sincronización entre el nodo de control y el resto de nodos. De la misma manera hemos establecido dos modos básicos de funcionamiento respecto a dónde se realizan los cálculos:

- **Modo 1: Cálculo en nodo de control.** En este modo los nodos receptores mandan las señales de audio grabadas al nodo de control, donde se realizan los cálculos necesarios para obtener los parámetros acústicos. El nodo de control se encargará de almacenar los parámetros calculados y subirlos a la nube para su visualización y consulta o análisis.

- **Modo 2: Cálculo en nodos receptores.** En este modo los nodos receptores se encargan de grabar las señales de audio y además de realizar los cálculos necesarios para obtener los parámetros acústicos deseados. Una vez realizados los cálculos, lo único que se transmite es el valor de los parámetros acústicos calculados al nodo central, que se encargará de almacenarlos y subirlos a la nube como en el modo anterior.

En ambos modos de funcionamiento estamos hablando de computación en la frontera o *Edge Computing*, ya que realizamos todos los cálculos de manera local y empleamos la nube únicamente para almacenar los valores calculados de los parámetros acústicos, con lo que

ahorramos ancho de banda descargando la red y mejorando tiempos de respuesta a la hora de enviar los datos.

Llegados a este punto cabe mencionar que el nodo de control puede estar implementado por cualquier dispositivo que disponga de ciertas características básicas como son la capacidad de emitir una red wifi o conectarse a un dispositivo que lo haga, (enrutador o router wifi), capacidad de almacenamiento de datos y cierta capacidad de cálculo. De manera paralela ocurre con los nodos receptores, que dado el diseño realizado del sistema, pueden estar implementados en cualquier dispositivo que permita grabar señales acústicas y posea cierta capacidad de cálculo, además claro está, de la conectividad inalámbrica necesaria. Esta propiedad nos ha permitido ensayar con diferentes dispositivos tanto en los nodos receptores como, sobretodo, en el nodo de control, evaluando el desempeño a la hora de realizar los cálculos necesarios para obtener los parámetros acústicos en cada caso, como se describirá con detalle en una sección posterior.

En nuestro caso hemos empleado 2 tipos principales de dispositivos: ordenadores personales o PC de sobremesa y portátil y ordenadores de una sola placa o SBC, de *Single Board Computers*. La capacidad de cálculo en los SBC se ha incrementado exponencialmente en los últimos años, esto junto con su capacidad de conectividad, la posibilidad de trabajar con un conjunto de sensores muy amplio y el bajo consumo, los ha definido como el dispositivo idóneo para las redes IoT. Debido a la movilidad necesaria en los nodos receptores, en nuestro sistema hemos empleado sólo placas SBC para estos nodos, más en concreto, la Raspberry Pi 3B[18] , a la que se ha incorporado un micrófono omnidireccional de condensador, reducidas dimensiones y bajo coste, el Andoer B01LCIGY8U-USB[19] . Este micrófono se ha escogido por su buena relación calidad-precio, ya que para conseguir micrófonos de características notablemente superiores habría que multiplicar por 5 su coste. A los nodos receptores se ha incorporado un batería externa que permite varias horas de autonomía gracias al reducido consumo de los mismos. En la Figura 2.21 se puede ver un nodo receptor del sistema IoT diseñado.

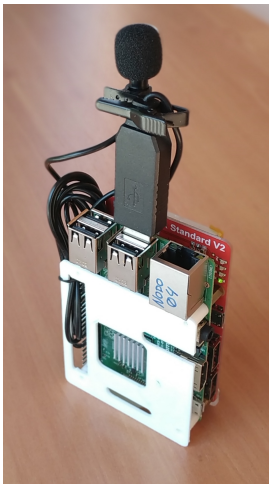


Figura 2.21: Nodo receptor del sistema IoT, basado en RPi3B y micrófono USB omnidireccional.

El caso del nodo de control es distinto, porque no requiere tanta movilidad, lo que ha permitido probar con todos los dispositivos listados y evaluar su desempeño a la hora de calcular los parámetros acústicos necesarios en cada caso. La red WASN diseñada se ha convertido en un sistema IoT también por la adopción del protocolo empleado para las comunicaciones

[18]https://www.raspberrypi.com/products/raspberry-pi-3-model-b/

[19]https://www.andoer.com/p-d7103.html

entre los nodos receptores y el nodo de control. El protocolo de comunicaciones empleado es MQTT, del inlés *Message Queue Telemetry Transport*, que fué diseñado por IBM para la comunicación *Machine To Machine* o M2M. Este protocolo es muy ligero, de código abierto y funciona sobre TCP/IP, que lo hace independiente del lenguaje de programación y dispositivo empleado. Además permite la ejecución en el protocolo seguro SSL/TSL garantizando que las comunicaciones están cifradas y son seguras, incorporando también acceso por usuario y contraseña y varios niveles de QoS (*Quality of Service*) para garantizar las comunicaciones.

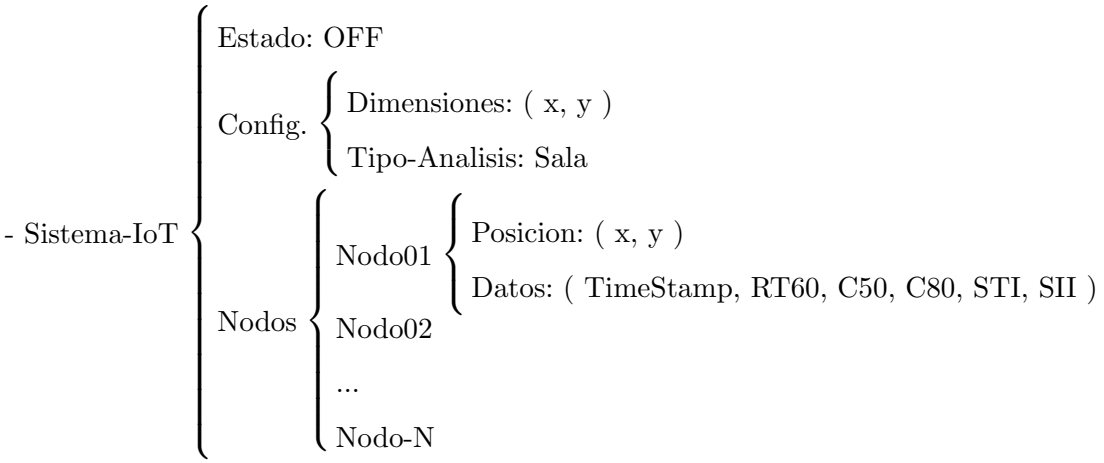


Figura 2.22: Ejemplo de topics definidos en el protocolo MQTT.

El funcionamiento de MQTT está basado en una estructura servidor-clientes, mediante la publicación y suscripción de mensajes, que gestiona el agente o *broker* MQTT. De esta manera, en el broker MQTT se definen unos temas o *Topics*, donde se puede publicar mensajes y a donde nos podemos suscribir para recibir los mensajes publicados. Así pues, se puede definir el tema "Sistema_IoT/" y los nodos pueden publicar mensajes añadiendo subtemas o etiquetas como por ejemplo: "Sistema-IoT/Nodos/Nodo01/Posicion..." o "Sistema-IoT/Nodos/Nodo02/Datos...", facilitando la escalabilidad del sistema, pues cualquier nodo con el permiso necesario podrá crear sub-temas dentro del sistema y publicar los datos pertinentes. En la Figura 2.22 se puede ver un ejemplo de los temas definidos en nuestro caso, que se usan para controlar, leer y almacenar los valores publicados por los nodos en los mismos, realizando el cálculo de los parámetros en los propios nodos en el ejemplo. En nuestro caso el propio broker MQTT implementado en el nodo de control está suscrito al tema principal, de manera que recibe todos los mensajes publicados por los nodos. Toda la información se puede guardar en cualquier formato (texto plano, csv, etc), y por supuesto crear una base de datos con la misma estructura jerárquica de registros o campos que los temas definidos en el broker MQTT, pudiendo almacenar datos tanto en el nodo de control como en la nube. La sincronización entre los nodos receptores y el nodo de control se realiza mediante el protocolo NTP (*Network Time Protocol*) y se incluyen marcas de tiempo en los datos que se publican desde los nodos receptores, tanto si son secuencias de audio, como si son los valores de los parámetros calculados.

En la Figura 2.23 se muestra un ejemplo de medida usando el sistema IoT diseñado en un espacio docente, en este caso analizando SII, con una fuente de sonido marcada con el rectángulo en la figura, simulando ser un orador y los nodos receptores marcados con círculos en las posiciones deseadas. Como el sistema es escalable, se ha probado hasta con 8 nodos receptores aunque finalmente establecimos en 4 los nodos receptores pudiendo emplearse más en caso de ser necesario. Dado el diseño del sistema es muy sencillo cambiar la posición de los nodos para tomar medidas en nuevos puntos tantas veces como se desee.
La única limitación detectada y que puede presentar el sistema diseñado tiene que ver con el

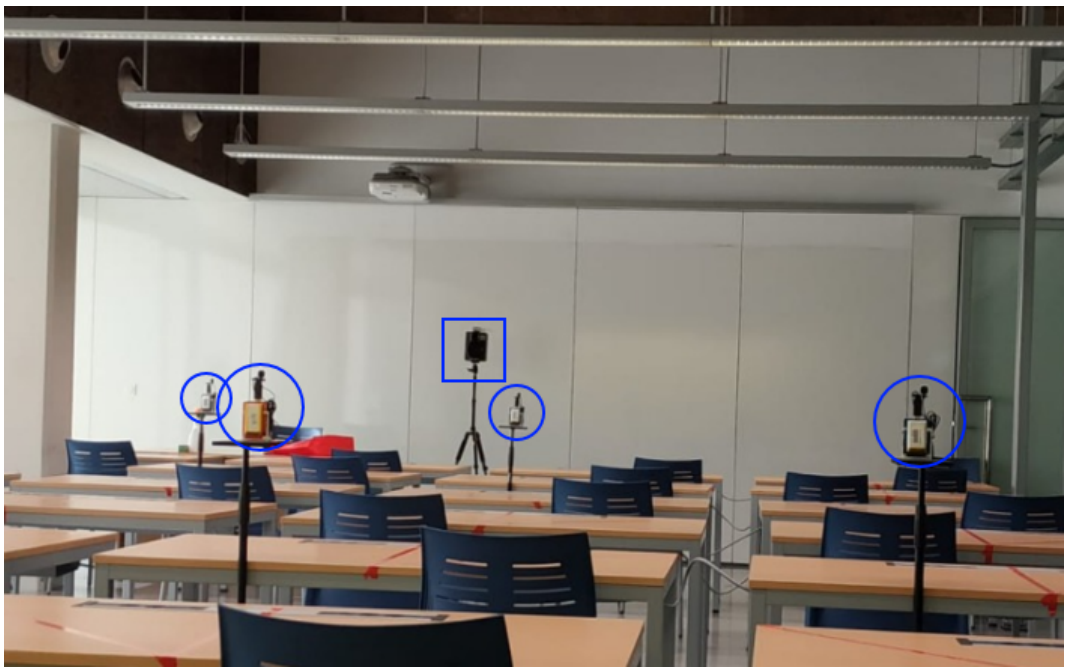


Figura 2.23: Ejemplo de medida en una sala docente con el sistema IoT diseñado.

tiempo de cálculo de los parámetros acústicos, cuando se desea realizar una monitorización continua durante horas, como por ejemplo al evaluar el conjunto de parámetros de molestia psicoacústica en una zona. En este caso, mediante el cálculo directo ninguno de los dispositivos informáticos testeados consiguen calcular los parámetros acústicos en tiempo real, como se verá más adelante. Es decir, para calcular los parámetros de molestia psicoacústica de una secuencia de audio de 1 segundo, el proceso de cálculo emplea más de 1 segundo, por lo que aunque se acumulen secuencias de audio en memoria, a modo de *buffer*, el tamaño de esta memoria intermedia irá creciendo de manera constante, lo que puede presentar un problema. Esto ocurre con el sistema funcionando en cualquiera de los 2 modos definidos, calculando en el nodo de control o calculando en los nodos, por lo que se ve claramente que el cuello de botella es la velocidad de cálculo del sistema, que es lo que se debe optimizar para permitir el funcionamiento en tiempo real, realizando además idealmente los cálculos en los nodos. No obstante esto no ha empañado el excelente funcionamiento del sistema, ya que hemos esquivado este cuello de botella de diversas maneras para permitir la monitorización continua de soundscapes: creando un buffer limitado de memoria que pausa la monitorización cuando llega a cierto umbral, para procesar las secuencias de audio almacenadas, monitorizando sólo un porcentaje de tiempo para permitir calcular, por ejemplo el 50 % del tiempo, o almacenar todas las secuencias de audio de una sesión de monitorización en disco duro, para realizar los cálculos a posteriori, aunque este es el peor caso, pues se trabaja a ciegas hasta que no se analizan los datos y correlar los eventos acústicos monitorizados se vuelve complejo.

Fuera de la monitorización continua durante intervalos de tiempo extensos, propia de los parámetros de molestia psicoacústica, no se ha detectado ningún problema. El conjunto de parámetros de sala no precisa estrictamente de monitorización continua para analizar el comportamiento acústico de un recinto, pues se analizan tantas veces como se quiera las señales reproducidas por el mismo sistema y se devuelve el valor de los parámetros calculados. Si bien hay casos en que es necesario observar cómo varían los parámetros acústicos, mientras se incrementa el número de personas presentes en una sala, por ejemplo. En cualquier caso resulta complejo y largo el proceso de extracción de la respuesta impulsiva de una sala, por lo que ser capaz de extraer los parámetros acústicos de sala a partir de señales acústicas más

sencillas o presentes en el propio recinto, es especialmente interesante. Si además se consiguiera realizar los cálculos en el propio nodo inalámbrico en un tiempo relativamente inferior al empleado con el método habitual, se lograría un avance muy significativo en lo que a sistemas de monitorización acústica se refiere, más teniendo en cuenta que los sistemas IoT se enfocan a monitorización continua.

Así pues, además del buen funcionamiento del sistema, se ha expuesto la importancia que tiene acelerar al máximo el cálculo de los parámetros para permitir el funcionamiento en tiempo real y sobretodo el cálculo de los mismos dentro de los nodos receptores, lo que permitiría obtener los datos de manera instantánea y monitorizar de forma continua. Por este motivo enfocamos nuestra atención a las redes neuronales convolucionales como herramienta matemática que nos permita acelerar el proceso de cálculo y poder realizarlo con cierto margen dentro de los nodos receptores.

Para concluir este apartado cabe mencionar que el sistema IoT descrito se ha empleado en la multitud de análisis acústicos realizados a lo largo de esta investigación, proceso mediante el que hemos llegado a su diseño final. Esto se ha plasmado en los artículos de congreso [79, 80] además de en los artículos de revista [24, 25, 26, 27], incluidos en el compendio de esta Tesis como Anexos A, B, C y D.

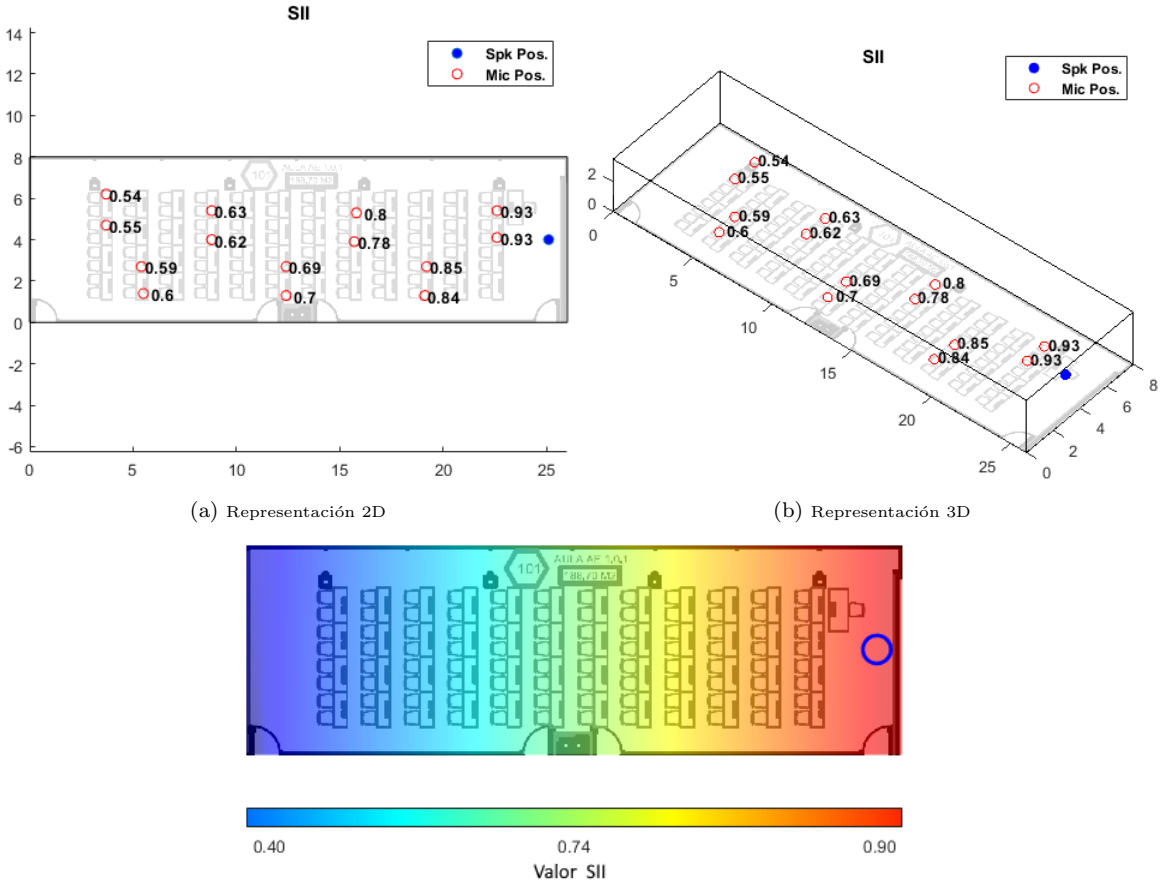
2.3. Sistema de Representación

Los dos conjuntos de parámetros acústicos que se han descrito nos permiten realizar un análisis riguroso de la molestia psicoacústica de un soundscape y del comportamiento acústico de un recinto. En ambos casos, las magnitudes calculadas tienen relación directa con el espacio físico estudiado y es esencial disponer de un sistema de representación que ayude a realizar un análisis acústico pormenorizado y facilite la representación de los datos obtenidos, más allá de listar los mismos como texto. En nuestro caso, una vez desarrollados los algoritmos y sistemas de medida descritos y que de por sí presentan una evolución respecto a los existentes, se planteó la necesidad de un sistema de representación acorde, que mejorara también las características de representación de los sistemas actuales, facilitando la comprensión y análisis de los parámetros acústicos calculados, tal como se describe en esta sección.

Tanto en el sistema de medida cableado como en el sistema IoT se puede acceder a los datos calculados desde un equipo informático estándar, por lo que hemos aprovechado su capacidad de cálculo y representación gráfica para mostrar la información de una manera muy intuitiva. De esta manera, se ha diseñado el sistema de representación empleando la plataforma Matlab 2019b[20] , que proporciona las herramientas de cálculo y representación necesarias. Así pues, el sistema de representación diseñado permite 3 tipos básicos de representación para cualquier parámetro acústico que se desee mostrar: Representación en 2 dimensiones, en 3 dimensiones y en forma de mapas de calor 2D.

En la Figura 2.24 se puede ver un ejemplo de los tres tipos básicos de representación, en este caso a la hora de analizar la inteligibilidad del habla o *SII* en el aula docente 101 de la Escuela Técnica Superior de Ingeniería de la Universidad de Valencia (ETSE-UV). En la representación en 2 y 3 dimensiones mostradas en las Figuras 2.24a y 2.24b respectivamente, se muestran los ejes en metros además de una imagen de fondo, si se dispone de ella, para facilitar el análisis del entorno. Esta imagen de fondo se inserta también en el mapa de calor para facilitar la ubicación de zonas, como se puede ver en la Figura 2.24c. Tanto

[20]https://es.mathworks.com/products/matlab.html



(a) Representación 2D

(b) Representación 3D

(c) Representación como mapa de calor.

Figura 2.24: Sistema de representación gráfica de parámetros acústicos. Valores de SII en el aula 101 de la ETSE-UV.

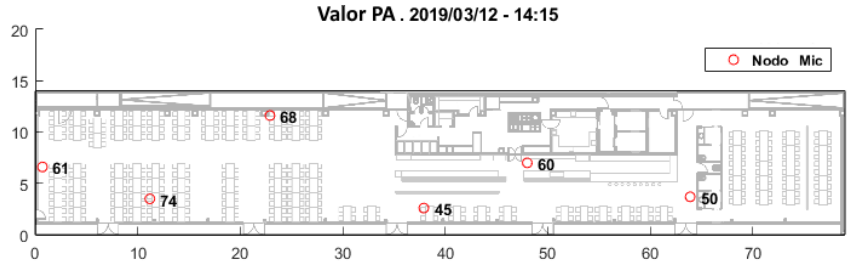
las posiciones de los nodos como las dimensiones de la zona analizada y el resto de opciones de representación son configurables por el usuario, como el hecho de mostrar posiciones de nodos y sus valores, leyenda, etc.

Nodo	Tiempo	N	S	F	R	PA
01	{'2019-04-10 17:05'}	14.84	1.33	2.34	0.11	26.07
02	{'2019-04-10 17:05'}	24.00	1.34	3.41	0.46	48.36
03	{'2019-04-10 17:05'}	20.71	1.33	2.15	0.16	33.98
04	{'2019-04-10 17:05'}	20.07	1.31	1.46	0.04	28.74
05	{'2019-04-10 17:05'}	10.48	1.37	1.67	0.01	16.65
06	{'2019-04-10 17:05'}	21.27	1.35	0.99	0.04	27.85

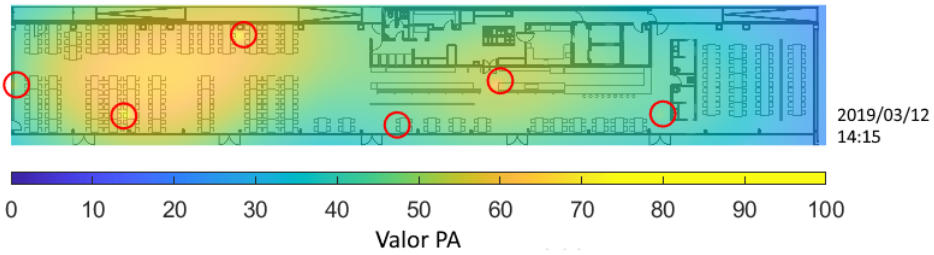
Figura 2.25: Representación numérica de parámetros acústicos.

En la Figura 2.25 podemos ver un ejemplo de la representación numérica de los parámetros de molestia psicoacústica para un análisis de 6 nodos en un momento concreto, que se puede emplear de manera complementaria a la parte gráfica para evaluar los datos de análisis. Se aprecia fácilmente que para muchos parámetros, el hecho de disponer de un sistema de representación gráfica avanzado nos ayuda a interpretar más fácilmente los datos obtenidos. Por ejemplo, en la Figura 2.24 antes, se representan los valores de inteligibilidad SII en cada

una de las 14 posiciones estudiadas del aula 101 de la ETSE-UV, pudiendo detectarse rápidamente las zonas conflictivas y actuar en consecuencia para mejorar la inteligibilidad del habla.



(a) Representación 2D



(b) Representación como mapa de calor

Figura 2.26: Representación de valores de PA en la cafetería de la ETSE-UV.

Similarmente en la Figura 2.26 se observa cómo varía el indicador general de molestia psicoacústica PA a lo largo de la cafetería de la misma Escuela de Ingeniería, en este caso para una hora concreta y donde se puede ver tanto los valores numéricos por nodo como el mapa de calor resultante de las 6 posiciones analizadas, de una manera mucho más intuitiva que observando la tabla mostrada en la Figura 2.25, de forma que podemos identificar rápidamente las posiciones de los nodos, los valores asociados y por lo tanto la molestia producida por el sonido presente en cada zona. Las interpolaciones de valores para realizar los mapas de calor se pueden configurar a voluntad y se basan en los métodos de creación de superficies suavizadas mediante interpolación lineal, polinomios de grado 3 e interpolación kriging, según la distribución del parámetro que se desee mostrar para permitir la mejor representación posible.

Para estudiar cómo varía un parámetro acústico a lo largo del tiempo, como suele ser el caso de los parámetros de molestia psicoacústica, se puede realizar una representación de diferentes instantes seleccionados del volumen de datos, para apreciar su evolución. En la Figura 2.27 se puede ver las medidas del indicador general de molestia psicoacústica PA en 5 instantes de un día para la cafetería de la ETSE-UV. Se aprecia rápidamente cómo cambia la actividad de la cafetería a lo largo de las horas, dependiendo de la afluencia de público y los sonidos presentes en la misma, evaluando PA medido y representado como mapa de calor. Se puede ver también cómo se desplaza la actividad, que en las primeras horas se encuentra en la zona central de la cafetería, donde se ubican las cafeteras y servicio de desayuno, hacia la zona de la izquierda a medio día, donde se encuentran las mesas del comedor del alumnado.

Los mapas de calor generados son exportables y se pueden aplicar a representaciones similares en otros sistemas. Este es el caso del estudio realizado en la cafetería de la ETSE-UV, destinado a evaluar la molestia psicoacústica en espacios públicos a lo largo de periodos de actividad. En él se ha hecho uso de los mapas de calor generados para recrear un escenario virtual en tres dimensiones y por el que se puede mover el usuario y que además evoluciona con el tiempo, mostrándose en tiempo real o a partir de datos almacenados la molestia

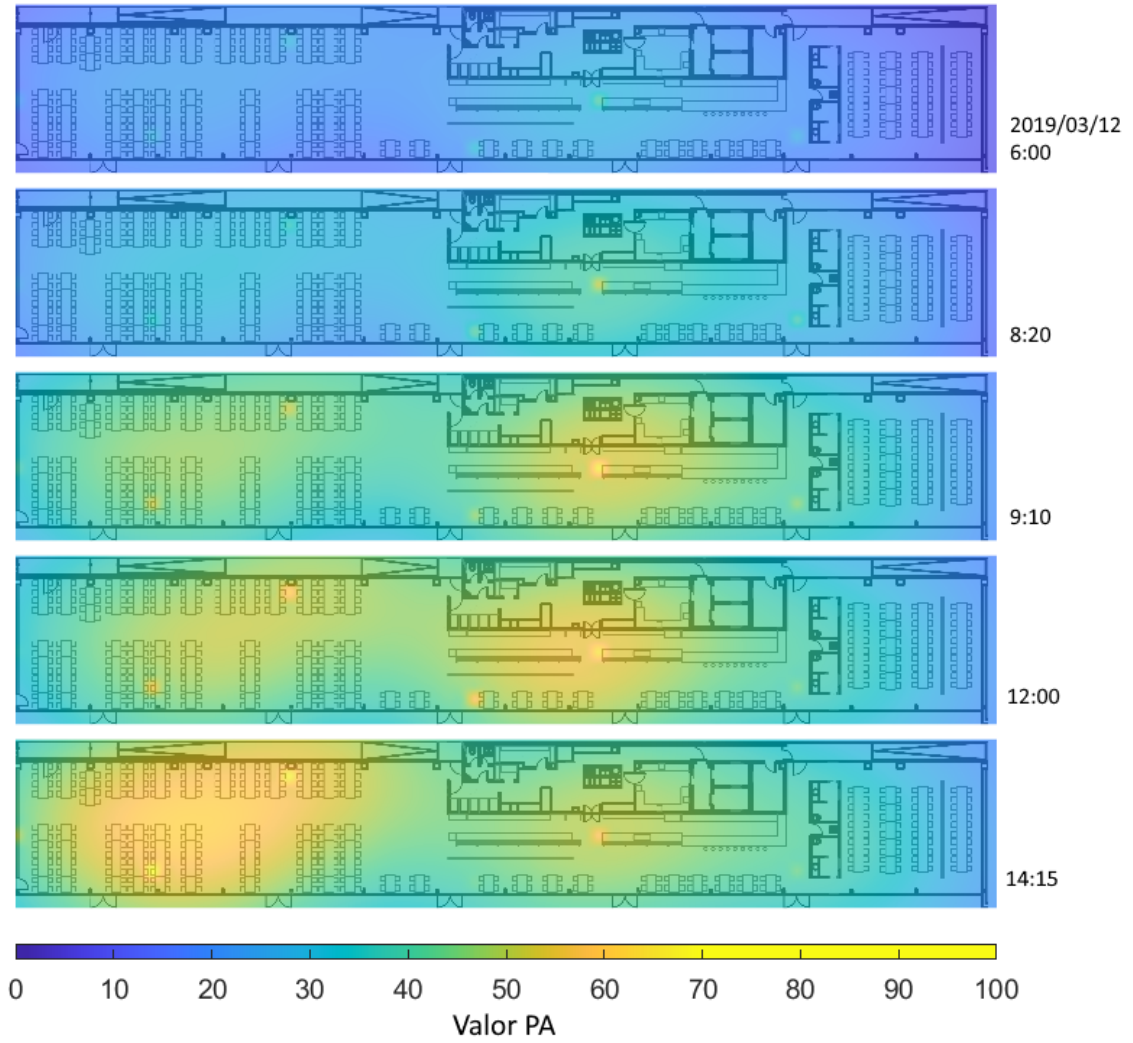


Figura 2.27: Evolución de PA a lo largo del tiempo en la cafetería de la ETSE-UV.

psicoacústica medida, como se puede ver en la Figura 2.28. Para realizar el modelado de la zona se ha empleado el software de diseño Autocad[21] (versión 2017) según los planos de edificación existentes y posteriormente se ha empleado la plataforma de desarrollo Unity[22] (versión 2020.315f1) para crear el escenario interactivo en 3 dimensiones e implementar el motor que permite el movimiento de un avatar mientras se integran los datos recibidos de molestia psicoacústica PA como mapa de calor semitransparente superpuesto a 1 metro de altura del suelo. Sobre la plataforma Unity se ha programado un script en lenguaje C# para recibir los datos del mapa de calor como objeto de tipo JSON (acrónimo de *JavaScript Object Notation*), un formato de texto plano muy extendido para el intercambio de datos independiente del lenguaje de programación empleado.

Tal y como se mencionaba, los datos se pueden mostrar en tiempo real mientras se monitoriza o en diferido a partir de los datos almacenados del período de actividad que nos interese. Pese a que esta última parte de la representación no es genérica, pues está creada a medida de un escenario concreto, sirve como ejemplo para remarcar las capacidades del sistema de representación diseñado, que permite de hecho presentar la información de una manera atractiva, intuitiva y clara para el usuario, no disponible en los sistemas comerciales

[21] https://www.autodesk.es/products/autocad

[22] https://unity.com/es

habituales.

El sistema de representación implementado aparece descrito en la comunicación de congreso *Visualization of nuisance information in acoustic environments using an IoT system* [81], y se ha empleado en los análisis acústicos incluidos tanto en los artículos de congreso [79, 80] como en los artículos de revista [24, 25, 26, 27], incluidos en el compendio de esta Tesis como Anexos A, B, C y D.

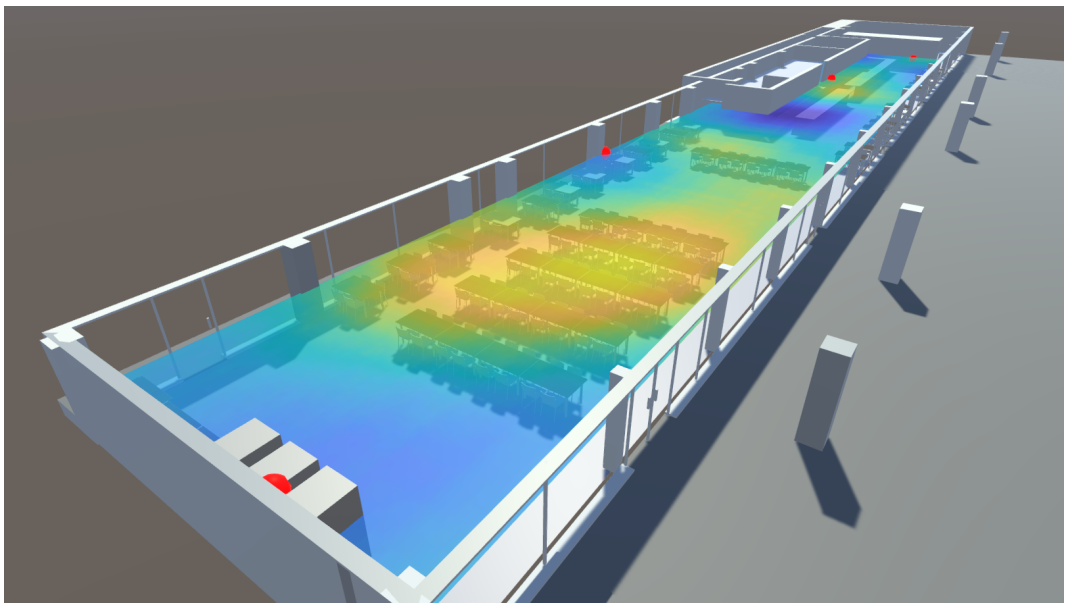


Figura 2.28: *PA* sobre un escenario 3D réplica de la cafetería de la ETSE-UV.

2.4. Cálculo directo y tiempo de procesado

A lo largo de este capítulo se han descrito dos conjuntos de parámetros acústicos, uno relacionado con la molestia psicoacústica y otro relacionado con parámetros de sala. Así mismo se han descrito también los dos sistemas de medida empleados, uno cableado empleado de forma clásica en investigación y otro más innovador basado en una red de sensores acústicos inalámbricos que hacen uso de tecnología IoT. En cualquier caso, los cálculos necesarios para obtener los parámetros acústicos se deben realizar o en un equipo informático habitual (PC), o en un dispositivo integrado tipo SBC (Raspberry Pi por ejemplo).

Concretamente, el sistema de medida basado en una red de sensores IoT, puede funcionar de dos maneras, calculando en el nodo de control o calculando en los nodos receptores de sonido. Se ha mencionado anteriormente que pueden presentarse problemas en ambos casos si se requiere una monitorización continua, debido a que el tiempo de cálculo requerido no permite el funcionamiento continuo en tiempo real. Es decir, se tarda más en calcular los parámetros, que la duración del fragmento de audio empleado para ello. En esta sección se describen con detalle las pruebas de rendimiento realizadas, empleando diferentes dispositivos para calcular los parámetros acústicos, además de los resultados obtenidos de las mismas que evidencian la característica mencionada.

En la Tabla 2.2 se resumen las características de los dispositivos empleados. Como puede verse, se han evaluado 2 tipos principales de dispositivos: ordenadores personales etiquetados como PC-1, PC-2 y PC-Portátil y ordenadores de una sola placa o SBC (de *Single Board*

Dispositivo	**CPU**	**Núcleos**	**Frec. (GHz)**	**RAM (GB)**	**Almacenamiento Tamaño / Tipo**
PC-1	Intel i7 7700	4	3.6	16	2 TB / HDD
PC-2	Intel Xeon 2× E5-2660v2	2×10	2.2	96	4 TB / HDD
PC-Portátil	Intel i7-1065	4	1.63	16	500 GB / SSD
UDOO X86 II-Ultra	Intel Pentium N3710	4	2.56	8	500 GB / HDD
Rarpberry Pi 3B	Broadcom BCM2837	4	1.4	1	16 GB / SDHC

Tabla 2.2: Dispositivos empleados en la evaluación de rendimiento.

Computers), como son la Raspberry Pi 3B y la UDOO X86. En ningún caso se ha empleado la unidad de proceso gráfico o GPU para realizar ningún cálculo en ningún dispositivo aunque se dispusiera de ella.

De esta manera el conjunto de dispositivos incluye una muestra de los más habituales a emplear en el sistema diseñado. La capacidad de cálculo en los SBC se ha incrementado exponencialmente en los últimos años, esto junto con su capacidad de conectividad, la posibilidad de trabajar con un conjunto de sensores muy amplio y el bajo consumo, los ha definido como el dispositivo idóneo para las redes IoT. El reducido coste económico de los mismos, comparado con las capacidades de conectividad, almacenamiento y cálculo que ofrecen, son un atractivo más para el uso de los dispositivos SBC.

2.4.1. Evaluación con parámetros de molestia psicoacústica

En la Figura 2.7 de la Sección 2.1.1 de esta Tesis se ha descrito el algoritmo diseñado para calcular los parámetros de molestia psicoacústica seleccionados, incluidos en el modelo de molestia de Zwicker. Resumiendo el proceso de cálculo, este parte de una señal de audio, se realiza un enventanado Hann para obtener *Loudness* (N) y *Sharpness* (S) y un enventanado Blackman para obtener *Roughness* (R) y *Fluctuation Strength* (F). Finalmente se emplean estos 4 parámetros para obtener el marcador de molestia psicoacústica general PA (*Psicoacoustic Annoyance*).

El análisis fundamental que se ha realizado ha sido la evaluación del tiempo de cálculo empleado para obtener los 5 parámetros mencionados, de manera que hemos establecido unas condiciones básicas para realizarlo. Como se ha descrito en la sección 2.1.1, en numerosos estudios donde ha participado el mismo Zwicker, para el caso de N y S, la señal de entrada se recorta con tamaños de ventana típicos de 80 a 200 ms, pero para el caso de R y F son necesarias ventanas más grandes (del orden de 500 ms mínimo), al tener en cuenta modulaciones de 4 a 15 Hz. Es por esto que hemos empleado ventanas de 1 segundo de duración como señal de audio de entrada. Como deseamos optimizar la velocidad de cálculo, se ha suprimido también el solape entre ventanas. Se ha considerado una frecuencia de muestreo de 16 KHz, que proporciona un buen equilibrio entre el número de muestras a procesar y el rango de frecuencias considerado por los parámetros evaluados.

Para este análisis del tiempo de cálculo se han empleado 1000 señales tomadas aleatoriamente de la base de datos de sonidos urbanos UrbanSound8K [82], que consiste en más de 8000 archivos de audio con sonidos típicamente urbanos adecuados al análisis de molestia psicoacústica a realizar, y que se describirá en profundidad en el siguiente capítulo. Estas 1000 señales se han remuestreado en los casos necesarios a 16 KHz y se ha extraído un segundo de un solo canal de cada una, para contar con 1000 señales de audio de 16000 muestras cada una. Tanto el tiempo de remuestreo, como las operaciones de lectura de disco no se han contabilizado en la medida del tiempo de cálculo para poder realizar las pruebas de manera objetiva en igualdad de condiciones iniciales, a partir del momento en que se tiene el audio cargado en memoria RAM.

Los dispositivos empleados en estas pruebas pertenecen al conjunto detallado en la Tabla 2.2 al principio de la Sección 2.4, de ellos hemos empleado los identificados como PC-1, PC-2 y Raspberry Pi 3B como dispositivo SBC. En la Tabla 2.3 se puede ver el tiempo empleado en calcular los parámetros psicoacústicos por tres dispositivos distintos, en este caso sin ninguna optimización de cálculo o procesado aplicada, siguiendo únicamente parámetros de la normativa y definición de cada caso. Se especifica tanto el tiempo empleado para calcular cada parámetro, como el tiempo total. El tiempo empleado en calcular PA una vez se tienen el resto de parámetros es del orden de 0.0002 segundos por lo que se ha incluido en la suma total del tiempo empleado. Como se observa, N y S son los parámetros más rápidos de calcular y R y F los que emplean más tiempo debido a las numerosas operaciones de filtrado adaptativo y a que los algoritmos en este punto no se habían desarrollado pensando en su eficiencia en tiempo de cómputo, sino en ajustarse a su definición teórica únicamente. Como consecuencia el tiempo total empleado en cualquier dispositivo supera ampliamente el segundo que dura una secuencia de audio de entrada. Una vez realizada esta primera serie de pruebas, y a la vista de los resultados, el resumen es positivo, al compararlos con cualquier sistema de análisis cableado, pues proporciona 5 parámetros acústicos en menos de 6 segundos en el peor de los casos, empleando sistemas comerciales de coste moderado o muy moderado. No obstante, a la hora de implementar un sistema de monitorización inalámbrico IoT, puede plantear los problemas que ya se han descrito, pues es necesario detener la monitorización mientras se calcula para no acabar desbordando la memoria, o acumular secuencias de audio durante cierto periodo (sin realizar cálculos), para proceder a realizar los cálculos más tarde sin monitorizar, lo que impide la monitorización continua en tiempo real.

Tabla 2.3: Parámetros psicoacústicos. Tiempos de cálculo directo (s), sin optimización.

Dispositivo	**N**	**S**	**R**	**F**	**PA (total)**
PC-1	0.1104	0.0001	0.6377	0.7291	1.4773
PC-2	0.1250	0.0001	0.7022	1.0128	1.8401
Rarpberry Pi 3B	0.2086	0.0004	2.9753	2.5199	5.6642

Debido a esto se realizaron varias optimizaciones de código íntegramente en Matlab para intentar atajar estos problemas. La primera de esta serie de optimizaciones se ha orientado a reducir las operaciones de lectura y escritura en disco que pueden ser muy lentas en el caso de discos duros físicos y de tarjetas de memoria en las SBC. La siguiente optimización consiste en realizar el pre-cálculo de los filtros en el caso de filtros no adaptativos que no dependan del contenido de la señal de entrada, para poder cargarlos en memoria RAM una única vez. De la misma manera, la última operación de optimización de esta etapa ha consistido en

almacenar toda esta información junto con los coeficientes y constantes empleados en uno o varios parámetros psicoacústicos en una archivo de inicialización que se carga en RAM una sola vez al inicio del proceso de cálculo. Los resultados de esta optimización se pueden ver en la Tabla 2.4.

Tabla 2.4: Parámetros psicoacústicos. Tiempos de cálculo directo (s), primera optimización.

Dispositivo	**N**	**S**	**R**	**F**	**PA (total)**
PC-1	0.0561	0.0001	0.5152	0.6429	1.2143
PC-2	0.0579	0.0001	0.5287	0.6511	1.2379
Rarpberry Pi 3B	0.0610	0.0001	0.8520	0.7423	1.6554

Como puede observarse la reducción del tiempo de cálculo es considerable, sobretodo en la Raspberry Pi, donde las operaciones de lectura de disco son críticas. Los resultados plasmados se incluyeron en la primera publicación del compendio de publicaciones incluida como Anexo A a la presente Tesis Doctoral, *Computation of Psycho-Acoustic Annoyance Using Deep Neural Networks* [24].

Los resultados obtenidos con código Matlab son muy prometedores pero los tiempos siguen superando el segundo, lo que puede seguir provocando problemas en la monitorización continua. Debido a esto continuamos optimizando el código del cálculo directo de los parámetros hasta llegar a tiempos inferiores al segundo finalmente. Para ello optamos por una parte en Matlab por revisar y precalcular cualquier dato posible para hacerlos comunes a todos los parámetros (sobretodo a *Rougness* y *Fluctuation Strength*, llevando al dominio frecuancial cualquier cálculo posible para ahorrar también operaciones en el dominio temporal, llegando por fin a bajar el tiempo total de cómputo hasta los 0.699 segundos para el PC-1 a costa eso sí de almacenar más información en memoria RAM, cosa por otra parte que representa un incremento insignificante en el uso de la misma dadas las capacidades del equipo. Por otra parte, aplicamos la misma filosofía pero convirtiendo el código a una combinación de Puthon que emplea librerías matemáticas programadas en C++. A diferencia de Matlab, que es un lenguaje interpretado como Python, C++ es un lenguaje compilado que se integra directamente con el lenguaje ensamblador del procesador, permitiendo una ejecución muy rápida del mismo. De esta manera, se han programado las funciones principales de cálculo en C++ siendo llamadas desde Python para lograr la mejor eficiencia. Además el compilador de C++ permite un porcentaje de optimización de código, que implementa una pseudo-paralelización del que lo permite, elemento que hemos aplicado para aumentar aun más la eficiencia en los casos en los que ha sido posible, como se muestra en la Tabla 2.5.

Se puede ver en la citada Tabla 2.5 como la aplicación de optimización de código en el compilador C++ ayuda a reducir los tiempos de ejecución tanto en el PC-1 como en la Raspberry Pi, sin embargo en esta última el incremento de rendimiento no es tan destacable como en el caso del equipo informático de sobremesa. Por añadidura la aplicación de esta optimización en el compilador lleva al procesador a trabajar a un porcentaje mucho más elevado por núcleo en la Raspberry, provocando que su temperatura permanezca de forma casi constante en 80^{o}C.

Cabe mencionar que debido a que se está grabando sonido en la propia Raspberry Pi mediante un micrófono USB, no se puede emplear refrigeración activa mediante un ventilador, por ejemplo, y únicamente se cuenta con refrigeración pasiva mediante disipadores de aluminio adheridos a los circuitos integrados de la misma. Se puede señalar que funcionando a

Tabla 2.5: Parámetros psicoacústicos. Tiempos de cálculo directo (s), segunda optimización.

Dispositivo (Lenguaje)	**N**	**S**	**R**	**F**	**PA (total)**
PC-1 (Matlab)	0.0375	0.0001	0.2846	0.3769	0.6991
PC-1 (C++/Python)	0.0810	0.0000	0.2979	0.2398	0.6187
PC-1 (C++/Python) Optimización de Compilador	0.0031	0.0000	0.1276	0.23499	0.3657
Rarpberry Pi 3B (C++/Python) Optimización de Compilador	0.0182	0.0001	0.7806	0.6809	1.4798

temperaturas tan elevadas en algunas ocasiones la Raspberry Pi ha entra en modo protección y reduce la frecuencia de reloj para protegerse hasta reducir la temperatura lo que aumenta el tiempo de cálculo reduciendo el beneficio obtenido. Esta última optimización que incluye elementos externos a Matlab ha servido como prueba de concepto, pero no ha sustituido el cálculo mediante Matlab, debido a los problemas mencionados ya que el ahorro de tiempo no es tan significativo como el incremento de complejidad del sistema que plantea. Además, a lo largo de las tablas presentadas se deduce que la velocidad de cálculo depende más de la frecuencia de funcionamiento de la CPU que del número de núcleos del procesador, pues la mayoría de operaciones no se pueden paralelizar y se alcanza una especie de techo práctico. Este efecto se percibe rápidamente entre los equipos PC-1 y PC-2, donde la frecuencia más alta de la CPU del PC-1 que puede llegar a los 4.0 GHz le proporciona una clara ventaja frente a los 2.4 GHz de la CPU del PC-2, pese a contar con 4 núcleos frente a los 20 de este último.

Como las optimizaciones de código que se han descrito pasan por cargar más elementos en la memoria RAM del dispositivo, el siguiente estudio realizado ha sido la evaluación del uso de RAM empleada. Sin embargo, como se puede ver en la Figura 2.6, el incremento de uso de memoria no es significativo comparado con el ahorro de tiempo que proporcionan. Así pues, empleando el código más optimizado de Matlab, que proporciona buenos resultados en tiempo de cálculo, el uso máximo de RAM es de unos 84 MB, que es una cantidad ínfima comparada con los tamaños de RAM evaluados, que parten en el peor caso desde 1 GB con la Raspberry Pi 3B.

Tabla 2.6: Parámetros psicoacústicos. Uso de RAM en calculo directo.

Dispositivo (Optimización)	**RAM (MB)**
PC-1 (No opt.)	26.3
PC-1 (Opt.)	83.8
Rarpberry Pi 3B (Opt.)	83.3

De cualquier manera, los excelentes resultados obtenidos en este proceso de optimización se han publicado también en el artículo *Psychoacoustic Annoyance Implementation With Wireless Acoustic Sensor Networks for Monitoring in Smart Cities* [83] ya que representa el logro de poder trabajar en tiempo real, pues empleando únicamente lenguaje Matlab ya se consigue bajar del segundo de tiempo en ejecución, al menos para el PC-1. En la Raspberry Pi no obstante ni siquiera aplicando todas las optimizaciones mencionadas se consigue bajar de 1.4 s de tiempo para calcular los parámetros psicoacústicos de un fragmento de 1 s de dura-

ción lo que como hemos visto implica inconvenientes en monitorización continua mediante el sistema IoT. Estos aspectos sobre el cálculo directo dentro del nodo se han publicado también en el artículo de congreso *Zwicker's Annoyance model implementation in a WASN node* [84], así como en los artículos de revista *Computation of Psycho-Acoustic Annoyance Using Deep Neural Networks* [24] y *Enabling Real-Time Computation of Psycho-Acoustic Parameters in Acoustic Sensors Using Convolutional Neural Networks* [25], incluidos en el compendio de publicaciones de esta Tesis.

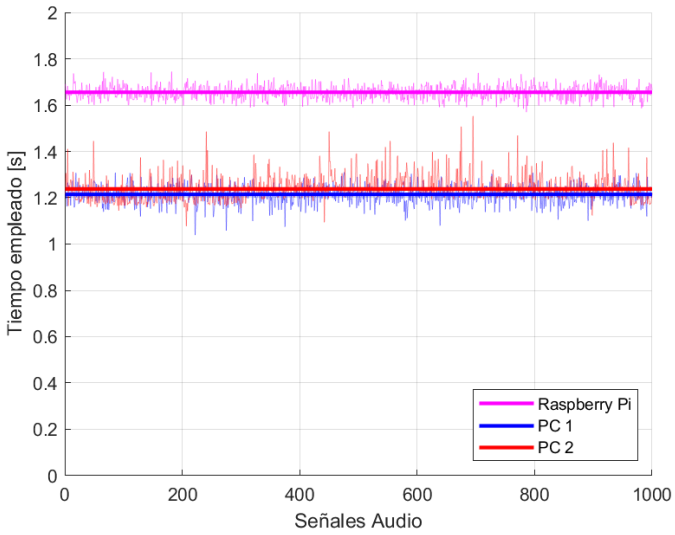


Figura 2.29: Parámetros psicoacústicos. Tiempos de calculo directo, para 1000 señales de entrada.

En la Figura 2.29 se puede apreciar también cómo, pese a haber optimizado el código, el tiempo de cálculo presenta una varianza elevada representada por picos que se alejan del valor medio. Esto es debido a que el procesador está ocupado por múltiples tareas, muchas de ellas ejecutándose en segundo plano, además del proceso de cálculo de los parámetros psicoacústicos, lo que plantea un problema de incertidumbre y provoca que necesitemos más margen que el obtenido con las optimizaciones del código de cálculo directo descritas. Ello ha propiciado la investigación en el uso de las redes neuronales convolucionales y al desarrollo de un modelo matemático de predicción de los parámetros psicoacústicos, como se describe más adelante en el capítulo 3.

2.4.2. Evaluación con parámetros de sala

El algoritmo diseñado para calcular el conjunto de parámetros acústicos de sala se ha mostrado en la Figura 2.16 de la Sección 2.1.2 de esta Tesis. Tal y como se ha descrito, los parámetros acústicos de sala caracterizan el comportamiento acústico de un recinto. Este comportamiento acústico depende de la geometría y los materiales que constituyen el espacio estudiado y no suelen variar si no se modifican. Por lo tanto una monitorización continua en el caso de los parámetros de sala no es tan esencial como el hecho de poder realizar un análisis de forma rápida y sencilla. El conjunto seleccionado de parámetros de sala comprenden el Tiempo de Reverberación ($RT60$), Claridad de la voz ($C50$), Claridad Musical ($C80$), *Speech Transmission Index* (STI) y *Speech Inteligibility Index* (SII). Los 3 primeros parámetros no deberían variar demasiado con la posición de estudio en la sala, pero en el caso de STI y SII esto cambia y la presencia de estructuras, mobiliario o la distancia pueden afectar mucho a la transmisión del habla y su inteligibilidad. Por ello su estudio en diferentes posiciones es

especialmente interesante y más aún si se desarrolla mediante un sistema IoT que facilite el proceso de medida y cálculo.

De modo similar a las pruebas realizadas con los parámetros de molestia psicoacústica, lo primero que se ha evaluado es el tiempo de cálculo de los 5 parámetros acústicos de sala. En este caso el procedimiento es mucho más complejo pues interviene el proceso de emisión de las señales, por lo que esta parte del procedimiento de medida no se ha incluido en los tiempos de cálculo. En consecuencia, la monitorización del proceso de cálculo comienza cuando se tienen en el sistema informático la señal MLS y la señal de habla recogidas por el micrófono emplazado en una posición concreta o de un nodo. Así, el proceso parte de la recepción de 2 señales acústicas, una proveniente de una señal MLS o un barrido en frecuencia y otra de una señal de habla. La duración de estas señales puede ser escogida por el usuario, pero para proceder a la evaluación de cálculo directo en igualdad de condiciones para todos los parámetros, hemos establecido la duración de los barridos en 20 segundos (siguiendo indicaciones de la norma ISO-3382) y la de la señal de voz en 3.5 segundos. Seguidamente en el algoritmo diseñado se realizan unos cálculos previos que incluyen la obtención de la respuesta impulsiva integrada de la sala mediante la señal MLS registrada. Fuera del cómputo, pues se realiza una única vez, se realiza la definición y creación en su caso de los filtros por bandas a emplear y la lectura de disco y carga de coeficientes constantes, de manera similar a las optimizaciones realizadas en el cálculo de los parámetros psicoacústicos que lograron excelentes resultados. Se procede a calcular entonces $RT60$, filtrar por bandas de 1 octava y calcular $C50$, $C80$ y STI. SII se calcula a partir de la señal de habla filtrada por bandas de 1/3 de octava. En esta evaluación se han empleado 1000 señales de audio de cada tipo, MLS y de hable. Las señales de habla empleadas pertenecen al idioma inglés en nuestro caso, pues hemos definido así el cálculo de SII, y proceden de la base de datos *DARPA-TIMIT Acoustic-Phonetic Continuous Speech Corpus*[85] que recoge señales de habla de 630 hablantes de diferentes sexos y dialectos.

En un análisis normal emplearíamos frecuencias de muestreo de 44100 Hz como mínimo para estudiar un rango más amplio de frecuencias, ya que el sistema de medida diseñado tanto cableado como IoT, lo permite fácilmente. No obstante, observamos que los parámetros $C50$ y $C80$ sólo contemplan para su cálculo bandas entre 500 Hz y 4 kHz. De la misma manera, en el cálculo de STI y SII se emplean sólo bandas entre 125 Hz y 8 kHz. Como $RT60$ evalúa la señal en el dominio temporal, la pérdida de contenido frecuencial, puede ser aceptable al menos para esta prueba concreta. Por ello a la hora de realizar el estudio del tiempo de cálculo directo, nos planteamos reducir la frecuencia de muestreo a 16 kHz. Los archivos de audio de la base de datos de habla están muestreados también a 16 kHz, con una duración media de 2 a 4 segundos, lo que ha propiciado establecer la frecuencia de muestreo en 16 kHz para la señal de habla a analizar, manteniendo coherencia también con el estudio realizado en el apartado anterior sobre los parámetros de molestia psicioacústica. Se logra así un buen equilibrio entre el número de muestras a procesar y el rango de frecuencias considerado para el habla y el cálculo del índice de inteligibilidad SII.

Tabla 2.7: Parámetros de sala. Tiempos de calculo directo (s).

Dispositivo	**C. Prev.**	**RT60**	**C50**	**C80**	**STI**	**SII**	**Total**
PC-1	3.4213	0.0017	0.0092	0.0091	0.1084	0.0878	3.6375
PC-Portátil	4.1297	0.0021	0.0166	0.0114	0.1361	0.1103	4.4063
UDOO X86	16.2116	0.0041	0.0440	0.0435	0.5087	0.4201	17.2320

Los dispositivos empleados en estas pruebas pertenecen al conjunto detallado en la Tabla 2.2 al principio de la Sección 2.4, de donde hemos seleccionado los identificados como PC-1, PC-Portátil y UDOO X86 como dispositivo SBC. En la Tabla 2.7 se pueden ver los resultados del tiempo de cálculo para los diferentes dispositivos. En este caso la respuesta impulsiva de la sala, incluida como cálculos previos con la etiqueta "C. Prev" de la tabla, se obtiene mediante la aplicación *Impulse Response Measurer* incluida en la toolbox de Matlab R2019b *Audio Toolbox 2.1* .[23] Esto dificultó ejecutarla de forma remota sobre un dispositivo que no soportara Matlab de forma nativa, por lo que se restringió a 3 dispositivos, incluyendo únicamente la placa SBC UDOO X86 como prototipo de dispositivo a emplear en un nodo IoT. Se puede ver que la mayoría del tiempo de cálculo se emplea en la extracción de la respuesta impulsiva, debido principalmente al análisis exhaustivo que realiza la aplicación de Matlab, ya que calcula la respuesta impulsiva en el dominio temporal, y en el dominio frecuencial en magnitud y fase mediante una interfaz gráfica muy completa, pero que genera un consumo de tiempo excesivo para nuestro propósito proporcionándonos información que para este análisis no necesitamos. En consecuencia se ha realizado una optimización en este aspecto implementando y empleando las funciones de cálculo de la toolbox de matlab fuera de la aplicación, calculando únicamente la respuesta impulsiva temporal sin interfaz gráfica tampoco que participe del proceso. Junto con esta optimización se ha simplificado el cálculo de algunos coeficientes empleados en el cálculo de STI y SII, logrando así la optimización del cálculo mostrada en la Tabla 2.8. Esta optimización y sobretodo la independencia de la interfaz del *Impulse Response Measurer* de Matlab ha permitido incluir la Raspberry Pi en la lista de dispositivos evaluados, teniendo un elemento más en los equipos SBC destinados a los nodos del sistema IoT, donde el cálculo en el propio nodo es más importante.

Tabla 2.8: Parámetros de sala. Tiempos de calculo directo optimizado (s).

Dispositivo	**C. Prev.**	**RT60**	**C50**	**C80**	**STI**	**SII**	**Total**
PC-1	0.7702	0.0018	0.0090	0.0088	0.1073	0.0620	0.9591
PC-Portátil	0.9851	0.0023	0.0111	0.0110	0.1248	0.0709	1.2052
UDOO X86	3.6845	0.0086	0.0429	0.0421	0.5136	0.2978	4.5896
Rarpberry Pi	4.5756	0.0107	0.0535	0.0523	0.6376	0.3684	5.6981

La Tabla 2.8 refleja fielmente la complejidad del cálculo directo, centrado el tiempo de cómputo en la obtención de la respuesta impulsiva y las operaciones de filtrado realizadas sobre la misma en los cálculos previos a los parámetros, seguida en tiempo de cálculo por los parámetros más complejos de calcular, como son STI y SII. La pequeña diferencia observada entre $C50$ y $C80$ se debe a la mayor simplicidad del cálculo de este último, ya que emplea únicamente la media de 3 bandas sin aplicar ponderación a las mismas. En resumen, el tiempo de cálculo directo estudiado muestra una velocidad aceptable comparada con la complejidad del proceso y más aun si se enfrenta a sistemas comerciales. Estableciendo las limitaciones configuradas en esta prueba, se logra obtener los 5 parámetros acústicos de sala prácticamente en 1 segundo de tiempo de cálculo empleando un ordenador personal y en unos 5.6 segundos empleando una SBC como la Raspberry Pi. Esto implica que en unos 10 segundos de análisis se puede tener una una idea precisa del comportamiento acústico de una sala, por lo que la complejidad de las medidas se centran en la instalación del sistema, el cableado, etc. Esto hace mucho más interesante aun el sistema IoT de monitorización diseñado, pues consigue el análisis en más tiempo pero facilitando el proceso enormemente.

[23] https://es.mathworks.com/products/audio.html

Acelerar el proceso de cálculo en un nodo del sistema IoT toma en este momento un nuevo interés y suscita un nuevo elemento que se podría optimizar: la obtención de los parámetros acústicos de sala sin el cálculo previo de la respuesta impulsiva del recinto. Esta idea se basa en emplear una red neuronal convolucional para predecir los parámetros acústicos de sala a partir de una señal acústica que se esté produciendo en el recinto y que por lo tanto esté afectada por las características acústicas de este, sin necesidad de calcular la respuesta impulsiva previamente. Como se mencionará más adelante también, hemos descrito el cálculo directo de los parámetros acústicos de sala y el tiempo empleado para obtenerlos, en las publicaciones de congreso [79] y [80], así como en los artículos de revista [26] y [27], incluidos en el compendio de publicaciones de esta Tesis como Anexos C y D.

Capítulo 3

Redes Neuronales

En este capítulo se aborda la descripción de las redes neuronales y en concreto la aplicación que se realiza en la presente Tesis de las redes neuronales convolucionales, mediante aprendizaje profundo, a la predicción de los parámetros acústicos descritos hasta el momento. En la sección 3.1 se realiza una pequeña introducción a las redes neuronales en general y a las redes neuronales convolucionales en concreto. En la sección 3.2 se describe todo lo relacionado con el modelo de red neuronal diseñado para la predicción de los parámetros psicoacústicos, diseño, base de datos, características de entrenamiento, pruebas y resultados. De la misma manera en la sección 3.3 se describe el modelo de red neuronal diseñado para la predicción de los parámetros acústicos de sala, su diseño, base de datos empleada, características de entrenamiento, pruebas realizadas y resultados obtenidos.

3.1. Redes neuronales convolucionales

A finales del siglo XIX Santiago Ramón y Cajal descubrió que el sistema nervioso humano está compuesto por una red de células individuales pero interconectadas entre sí, enviando y recibiendo información, las neuronas, lo que constituye la base de la neurociencia moderna. Gracias a este descubrimiento, ya en la primera mitad del siglo XX toma forma la idea de las redes neuronales, basada en imitar el comportamiento de las neuronas en el cerebro humano. Las conexiones o sinapsis neuronales se refuerzan cada vez que se utilizan, lo que sienta las bases del aprendizaje y la forma en que se concibe la inteligencia artificial.

La base de las redes neuronales es el Perceptrón, llamado también neurona artificial y definido en 1958, que puede considerarse una red neuronal de una sola capa, como se ve en la Figura 3.1. Aquí X_n representan las entradas, w_n los pesos sinápticos a entrenar, Y la salida del Perceptrón y $f(x)$ la función de activación. La función más sencilla $f(x)$ es la binaria, mostrada en (3.1) donde tomará el valor 1 o 0 dependiendo de si el producto escalar $w \cdot x$ supera el umbral u que representa el grado de inhibición de la neurona. Una forma muy esquemática de describir el aprendizaje o entrenamiento que fija el valor de los pesos w es teniendo una serie de valores conocidos de la salida δ_i para una serie de valores conocidos de las entradas $X_n i$. De esta manera puedo calcular el error cometido (δ_i - Y_i) e ir modificando los pesos en diferentes iteraciones para minimizar el error cometido entre el valor deseado a la salida δ y el valor obtenido a la salida Y.

$$f(x) = \begin{cases} 1, & \text{si } (w \cdot x) - u > 0 \\ 0, & \text{otro caso} \end{cases} \tag{3.1}$$

Figura 3.1: Perceptrón o neurona artificial con varios canales de entrada y 1 salida.

De esta manera, la tasa de actualización de los pesos se define en (3.2), donde α tomará un valor entre 0 y 1 si se emplea tasa de aprendizaje o 1 si no se desea emplear tasa de aprendizaje en la actualización de los pesos. En este caso, el error cometido entre el valor estimado y el valor real (δ - Y) es lo que deseamos minimizar en cada iteración y se denomina función de coste.

$$(w)' = w(j) + \alpha(\delta - y)x(j) \tag{3.2}$$

Las agrupaciones de neuronas forman las capas de una red neuronal, que constan de 3 tipos diferentes de capas: las de entrada, las de salida y las capas ocultas como se puede ver en la Figura 3.2. La capa de entrada está compuesta por neuronas que reciben los datos del entorno o de entrada y la capa de salida la forman neuronas de devuelven la respuesta de la red neuronal. Las capas ocultas no están conectadas al entorno y proporcionan la libertad a la red neuronal para aprender las características necesarias para modelar la respuesta deseada.

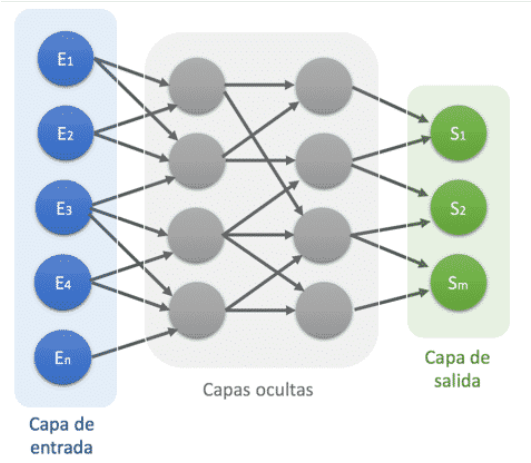


Figura 3.2: Esquema de red neuronal. Tipos básicos de capas.

Existen diferentes algoritmos de entrenamiento o aprendizaje que han llevado a definir 2 tipos principales:

- ***Machine learning*** o aprendizaje automático, comprende los algoritmos aprenden a partir de datos proporcionados la correspondencia entre las entradas y las salidas deseadas del sistema (si se le proporcionan), mediante parámetros ajustables cuyo valor se modifica en función de los datos disponibles, como por ejemplo se realizaría en un problema de clasificación. Generalmente los datos empleados en *Machine learning* requieren de un procesado previo para seleccionar los adecuados, hacer limpieza y alimentar el

sistema con algunas características concretas de los datos únicamente. Por ejemplo, empleando el histograma de color de las imágenes en vez de las imágenes completas, si los datos de entrada son imágenes. El aprendizaje automático no comprende únicamente a las redes neuronales y se emplea por ejemplo en máquinas de soporte vectorial [86], en arboles de decisión [87], algoritmos genéticos o algoritmos de reglas de asociación.

- ***Deep learning*** o aprendizaje profundo, comprende los algoritmos que intentan modelar abstracciones de alto nivel empleando redes neuronales con al menos una capa oculta, que admiten transformaciones no lineales e iterativas sobre datos que se expresan normalmente en forma matricial o como tensores [88]. En una cascada de capas se extraen y transforman variables, empleando la salida de cada capa como entrada de la siguiente y extrayendo las características de alto nivel derivándolas a un nivel inferior y formando una representación jerárquica. En el aprendizaje profundo no se suele extraer caracterísicas de los datos si no es estrictamente necesario y se suele alimentar la red neuronal con los datos en bruto, por ejemplo empleando la imagen completa sin procesar en vez de su histograma si hablamos del ejemplo mencionado. Como estructuras de aprendizaje profundo podemos encontrar las redes neuronales profundas o DNNs (*Deep Neural Networks*) con miles de capas multidimensionales generalmente no lineales y las redes neuronales convolucionales o CNNs (*Convolutional Neural Networks*) que se asemejan más a la corteza visual primaria del ser humano. Las CNNs consisten en múltiples capas de filtros convolucionales aplicados sobre matrices de una o más dimensiones y que son muy efectivas en tareas de visión artificial, segmentación y clasificación de datos gracias también a una alta velocidad de ejecución.

Tanto si hablamos de algoritmos de aprendizaje automático como de aprendizaje profundo, pueden clasificarse en tres tipos en función de los datos empleados para su entrenamiento:

- **Supervisado**, donde cada dato de entrada posee su etiqueta o dato de salida y el algoritmo debe crear una función que permita asignar a nuevos datos de entrada su correspondiente etiqueta o valor a la salida.
- **No supervisado**, donde a los datos de entrada no se les asigna una etiqueta clase o dato de salida y el algoritmo debe encontrar por sí mismo la similitud entre los datos.
- **Semi-supervisado**, donde algunos datos de entrada poseen datos de salida y otros no, empleándose técnicas mixtas entre ambos para resolver el problema.

El objetivo final del entrenamiento de una red neuronal es minimizar la función de coste (por ejemplo el error medio de estimación), como se ha mostrado en (3.2), encontrando los pesos adecuados para las capas de la red neuronal. El cálculo de estos pesos en un proceso de entrenamiento se lleva a cabo mediante un algoritmo llamado *backpropagation*, que se puede ver en la Figura 3.3.

Aquí aparece un elemento muy importante que es el Optimizador, ya que es el encargado de generar pesos mejores a cada iteración. La base del funcionamiento del Optimizador es calcular el gradiente de la función de coste (derivada parcial df/dw) por cada peso de la red. Como se desea minimizar el error, se modifica cada peso en la dirección negativa del gradiente (restando). Para agilizar la convergencia de la función de coste hacia el mínimo posible, se multiplica el vector gradiente por un factor llamado de entrenamiento, lr en la Figura 3.3. Al conjunto de métodos iterativos orientados a reducir la función de error al mínimo, se denominan métodos de optimización basados en gradiente descendente o de forma corta Optimizador.

En los últimos tiempos, las redes neuronales basadas en aprendizaje profundo han tomado especial protagonismo en la investigación en procesamiento de señales acústicas, abordando

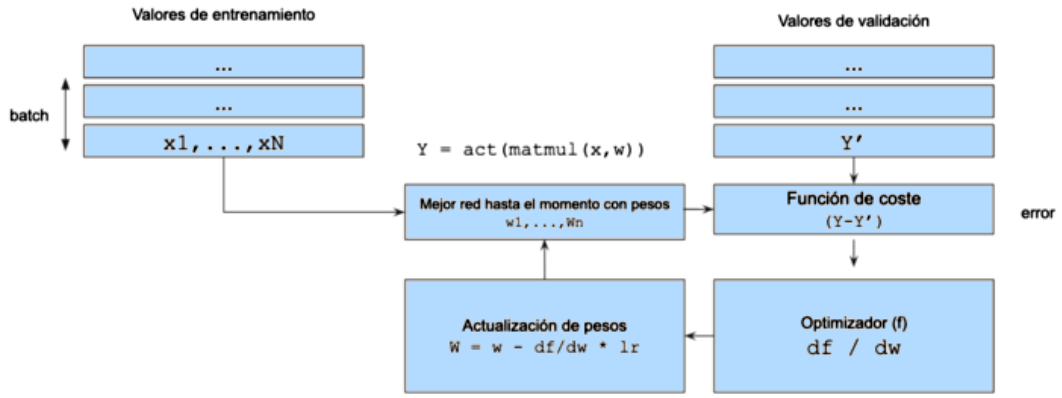


Figura 3.3: Algoritmo de *backpropagation*.

tareas como la separación y localización de fuentes sonoras [20, 17, 18] o la detección de eventos [19]. Las CNNs han sido empleadas en estos contextos con mucho éxito debido a su capacidad para aprender representaciones de características acústicas de manera optimizada. Tal y como se ha mencionado, es habitual encontrar CNNs dedicadas a clasificación de datos, por ejemplo a partir de vídeos, empleando la secuencia de imágenes o la señal de audio para clasificar lugares. Sin embargo, además de la clasificación, donde la salida toma una probabilidad asociada a cada clase definida, en otras aplicaciones la salida puede tomar la forma de uno o varios valores numéricos en función de la entrada presente. En estos casos se emplea una capa de regresión con activación lineal a la salida de la CNN, proporcionando una predicción del valor o valores que deseemos.

3.1.1. Pruebas preliminares y enfoque planteado

Vistos los distintos tipos de enfoques desde los que se pueden emplear las CNNs para solucionar nuestro problema de velocidad y complejidad de cálculo, en un principio contemplamos todo el abanico posible y fuimos descartando opciones mediante diferentes pruebas preliminares realizadas.

Tanto para los parámetros psicoacústicos, como para los parámetros acústicos de sala deseamos predecir de manera muy rápida una serie de valores a partir de la señal de audio presente en al entrada del sistema, empleando el mínimo de procesado posible de esta señal. Al principio de la investigación, consideramos enfocar el problema empleando como entrada a nuestro modelo de CNN representaciones basadas en características del sonido. Este es un planteamiento que se puede encontrar en diferentes publicaciones [21, 22, 23, 89, 90], generalmente enfocadas a la clasificación, donde se emplean espectrogramas, espectrogramas de MEL, Coeficientes Cepstrales en las Frecuencias de Mel (MFCC), o filtros gammatonales, por nombrar algunos ejemplos, como entrada a los modelos de CNN, por lo que nos pareció un punto de partida adecuado. Este enfoque se encuadraría dentro del aprendizaje automático o machine learning, no obstante lo descartamos conforme comenzamos a desarrollar la investigación, debido a la complejidad del procesado de señal previo requerido por la cantidad de características de entrada necesarias. En nuestro caso el tipo de caractreísticas extraidas del audio eran dependientes del parámetro acústico concreto a calcular, por lo que fuimos sumando características, debido a que pretendemos diseñar un modelo de CNN que sesa capaz de predecir un conjunto de parámetros de forma simultánea. Como consecuencia, llegamos al punto en que el procesado de señal previo para obtener el total de características superó el umbral límite para ser útil comparado con el cálculo directo de los parámetros acústicos en

Tabla 3.1: Ejemplo de escala empleada para dividir los valores de PA en 5 clases.

Molestia subjetiva	**Intervalo de PA**
Ninguna	0 - 3.6
Ligera	3.7 - 23.5
Moderada	23.6 - 58.6
Mucha	58.7 - 89.7
Extrema	89.8 - 100

cada caso. Debido a esto afrontamos un diseño basado en aprendizaje profundo o deep learning, donde el procesado de señal previo a la entrada del modelo de CNN es mínimo o nulo y donde es el propio modelo el que extrae las características útiles de la señal de entrada. Pese a que llegado un punto, no desarrollamos más el entrenamiento del modelo de CNN empleando la extracción de características, los primeros resultados respecto a precisión fueron también mejores empleando aprendizaje profundo y eso sumado a la idea de realizar el procesado en dispositivos de bajo coste con capacidades limitadas, acabó de inclinar la balanza.

Dentro de estos estudios preliminares, evaluamos también la forma de la respuesta de salida de la red neuronal. Pese a que pueda parecer la mejor solución optar por un modelo basado en regresión de salida, ya que deseamos obtener un valor numérico para cada parámetros acústico, inicialmente no despreciamos el enfoque de emplear una capa de clasificación de salida. Debido a que para implementar un problema clasificación, debemos de disponer de los datos de entrada divididos en diferentes clases, procedimos a dividir el rango completo de valores de cada parámetro acústico en diferentes clases. Es decir, por ejemplo para el caso de Loudness, dividimos el rango completo de valores disponibles en 5 intervalos equiespaciados, que representarían 5 clases distintas. Para el indicador general de molestia PA cuyos valores van de 0 a 100, establecimos una escala de 5 intervalos tanto lineales como de diferente tamaño, basados en estudios como los de J.M. Fields [91] o el de E. Hernández [92] por ejemplo, que observan la norma ISO/TS 15666:2021, *Acoustics — Assessment of noise annoyance by means of social and socio-acoustic surveys* [93] para definir la escala que se muestra en la Tabla 3.1 y que representan intervalos definidos mediante encuestas. Se probó también escalas con 10 clases en estos estudios preliminares, sin embargo los resultados fueron peores en estos ensayos que los obtenidos empleando regresión en la capa de salida.

De esta manera, los test preliminares han ayudado a definir tanto para los parámetros psicoacústicos como para los parámetros acústicos de sala un enfoque basado en regresión, empleando aprendizaje profundo supervisado, y por lo tanto partiendo de señales de audio sin procesar como entrada de dos CNNs (una para cada grupo de parámetros acústicos), que proporcionarán a la salida la predicción de los parámetros acústicos adecuados a cada caso.

3.2. CNN aplicada a parámetros psico-acústicos

La primera aplicación que hemos hecho de las redes neuronales convolucionales ha sido a la predicción del conjunto de parámetros psicoacústicos, para conseguir los valores de los mismos en un tiempo inferior al empleado por el cálculo directo. Recordemos que el tiempo requerido por el cálculo directo imposibilitaba el funcionamiento del sistema de monitorización IoT en tiempo real, o al menos dificultaba su uso a la hora de realizar monitorización continua.

El uso de una arquitectura CNN de extremo a extremo está motivado por su capacidad de aprender e implementar filtros centrados en diferentes bandas sobre la entrada, de manera similar a como se comportan los múltiples filtros por bandas que emplean los algoritmos de cálculo directo.

Para proceder al diseño y entrenamiento de la CNN el primer punto es contar con suficientes datos para llevarlo a cabo, por lo que establecer una base de datos bien definida es esencial.

3.2.1. Base de datos

Las secuencias de audio que forman la base de datos diseñada para nuestra CNN ha sido extraída de la base de datos UrbanSound8K [82]. Esta consiste en más de 8000 archivos de audio con sonidos urbanos que siguen una taxonomía respecto a los espacios a estudiar de una ciudad similar a la definida por la norma ISO 12903:2, *Acoustics-Soundscape - Part 2: Data collection and reporting requirements*[94].

Aunque esta sea una base de datos dedicada a clasificación, en nuestro caso no vamos emplear las etiquetas dado que no vamos a clasificar sonidos. Considerando todos los archivos con longitud superior al segundo, se ha procedido a dividir los más largos y obtener un total de 59000 segmentos de audio de 1 s de duración. A estas grabaciones se les ha sumado un total de 1150 secuencias más (de 1 s de duración también) grabadas en distintos lugares de diferentes ciudades y pueblos estudiados, llegando a un total de 60150 señales de audio. Los archivos originalmente están grabados con frecuencias de muestreo que van de los 8 kHz a los 48 kHz, por lo que se ha procedido a remuestrearlos a 16 kHz, que es la misma frecuencia empleada en el estudio de tiempo de cálculo realizado sobre el sistema de medida empleando cálculo directo. Como ya se ha descrito, esto nos proporciona una buena relación de compromiso entre el número de muestras a analizar y el rango de frecuencias a estudiar, pues en sonidos que producen una molestia psicoacústica elevada rara vez hay componentes frecuenciales que superen los 8 kHz. Respecto a la amplitud de las señales, en nuestro caso no se pueden normalizar, pues la amplitud de la señal es esencial para el cálculo de los parámetros psicoacústicos, por lo que se verán afectadas de cualquier efecto producido por las diferentes ganancias de los micrófonos y la distancia de las fuentes de sonido. Sin embargo esto no hace sino mantener el realismo y la integridad de los datos, que han sido tomados en escenarios reales donde en una situación habitual, al oyente que transite por una ciudad también le causarán más molestia ciertos sonidos cuanto más cercanos sean.

A continuación se han calculado mediante el algoritmo de cálculo directo los parámetros de molestia psicoacústica de cada señal de audio, además del indicador general de molestia: *Loudness* (N) y *Sharpness* (S), *Roughness* (R), *Fluctuation Strength* (F) y PA. Finalmente, antes de proceder al entrenamiento de la red neuronal, se ha dividido la base de datos entera de forma aleatoria en 3 conjuntos de datos:

1. **Entrenamiento**, con 47320 señales.

2. **Validación**, con 11830 señales.

3. **Test**, con 1000 señales.

Los conjuntos de datos que participan en el proceso de entrenamiento de la CNN son el de Entrenamiento y el de Validación. El conjunto de datos de Test es independiente del proceso de entrenamiento y nos proporciona una manera fiable e independiente de probar la capacidad de la CNN diseñada de generalizar en la predicción de parámetros sobre señales de audio externas al proceso de entrenamiento.

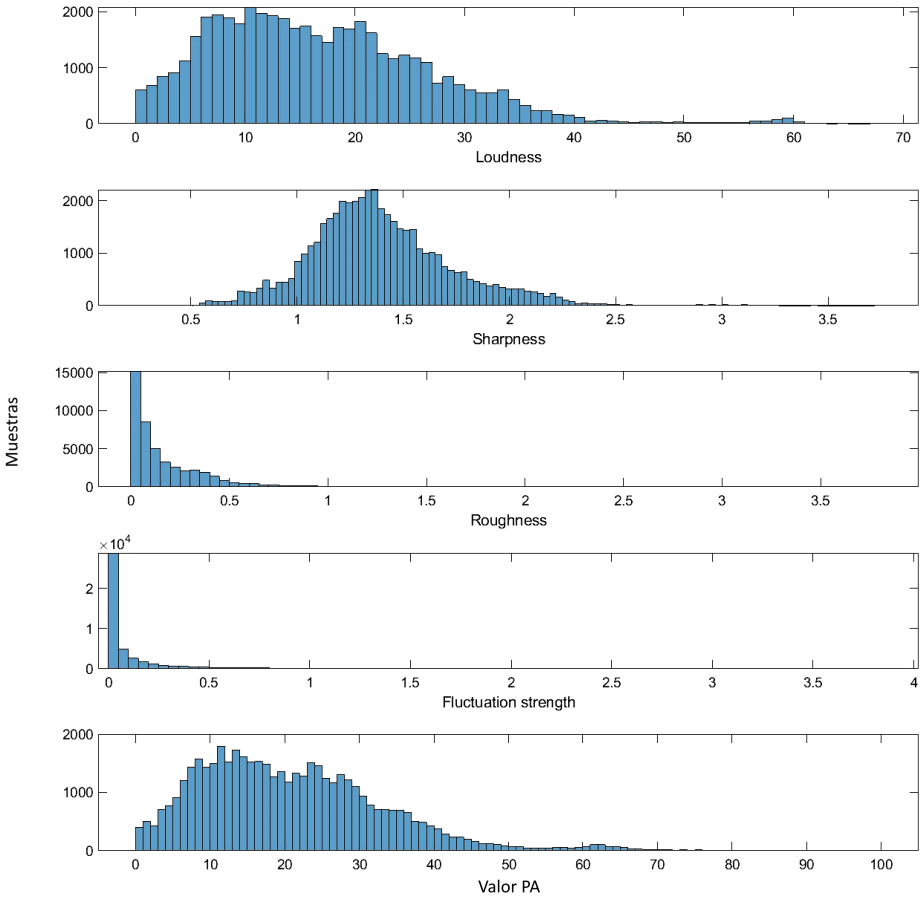


Figura 3.4: Histograma de valores calculados de N, S, R, F y PA en la base de datos creada.

En la Figura 3.4 se pueden ver los histogramas de los parámetros psicoacústicos calculados para todos los segmentos de audio de nuestra base de datos. Se observa que N es el parámetro que más influye en el indicador general de molestia psicoacústica PA, y que la mayoría de audios muestran valores muy bajos de R y F, ya que es difícil encontrar de forma natural en entornos urbanos normales sonidos modulados en frecuencias audibles. Esto se traduce en que las señales tienen un rango de valores reducido en estos parámetros y contribuyen a desequilibrar el conjunto de datos de entrenamiento. Al calcular PA a partir de sonidos urbanos reales no es habitual encontrar valores extremos que produzcan una molestia nula o extremadamente alta, recordando que PA se sitúa en el intervlo [0, 100]. Como la mayoría de archivos de audio poseen un PA situado en el intervalo de 0 a 50, esperamos tener más precisión en la predicción en este rango, así como esperamos menor precisión al predecir R y F debido al rango limitado de valores que contienen. No obstante hemos decidido no alterar la base de datos mediante variaciones de la misma empleando técnicas de aumento de datos, al menos antes de evaluar el desempeño de la misma, pues nos sirve como prueba de concepto a la hora de enfrentarnos al entrenamiento de una CNN con datos provenientes del mundo real naturalmente sesgados, lo que pondrá a prueba la capacidad de generalización de nuestra CNN. Además, en el caso de que nuestra CNN tenga un desempeño mejor en el rango de valores donde hay más señales de entrenamiento, si se evalúa en el mundo real tendrá un buen desempeño también, pues la base de datos es un fiel reflejo de este.

3.2.2. Diseño, configuración y entrenamiento

Una vez definidos los datos con los que podemos entrenar nuestra CNN, queda implementar el diseño de la misa, definiendo primero la estructura de capas que la van a formar y después las condiciones de entrenamiento de la misma, como se verá a continuación.

Para afrontar el diseño de nuestra CNN, se han analizado otras aplicaciones conocidas de las CNN, sobretodo en clasificación de imágenes, como son el diseño de la CNN AlexNet que dio lugar al proyecto ImageNet [95] para clasificación de imágenes, o la CNN SoundNet [18], empleada para identificar representaciones sonoras a partir de vídeos sin etiquetar. En estos enfoques de aprendizaje profundo el tamaño de los filtros disminuye desde la entrada hasta la salida, mientras que, de manera inversa, su número aumenta.

En este contexto, la selección de los hiperparámetros en nuestro sistema se seleccionó probando diferentes arquitecturas, que ajustamos de manera precisa hasta conseguir un rendimiento razonable. A lo largo de diversas pruebas establecimos un esquema básico para las capas convolucionales que proporciona un rendimiento óptimo en nuestra aplicación. Este esquema se basa en que las capas convolucionales no se emplean de manera aislada, sino que se incluyen en una unidad convolucional formada por 4 capas de dimensiones configurables. Esta unidad convolucional tiene la forma: Convolucional + Capa ReLU + Capa de Normalización + Capa MaxPool. Es decir, está formada por un bloque de convolución temporal seguido por una capa de activación *Rectified Linear Unit* (ReLU) que pone a 0 cualquier entrada negativa. A continuación hemos emplazado una capa de normalización por lotes que normaliza cada canal de entrada en un mini lote para acelerar el proceso de entrenamiento. Por último hemos incluido una capa de downsampling o diezmado (MaxPool) que guarda únicamente el máximo de los elementos indicados.

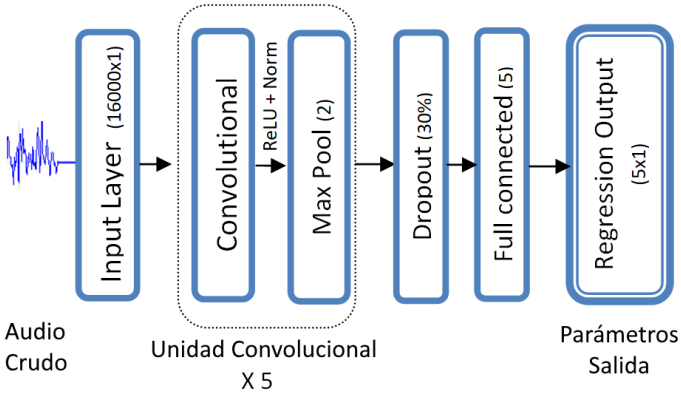


Figura 3.5: Diseño CNN parámetros psicoacústicos.

El diseño óptimo de nuestra CNN se puede ver de forma esquemática en la Figura 3.5 y consiste en 5 unidades convolucionales como la descrita, seguidas por una capa DropOut con una probabilidad de abandono de 0.3 que se sitúa detrás de las capas convolucionales para evitar el sobre-ajuste y por una capa totalmente conectada o FullConnected que reduce la salida a 5 elementos, llegando por fin a una capa de activación por regresión lineal de salida que proporciona la predicción de los 5 parámetros psicoacústicos descritos: N, S, R, F y PA.

La descripción completa de los parámetros correspondientes a cada capa de la CNN se muestra en la Tabla 3.2. Aquí se puede ver cómo partiendo de una señal de audio de 16000 muestras, a lo largo de la CNN el número de filtros convolucionales aumenta al mismo tiempo que disminuye su tamaño hasta converger en la capa de salida con las 5 predicciones de los

parámetros. La profundidad de la red seleccionada finalmente, que ha proporcionado el mejor comportamiento en predicción, es considerablemente menor que la utilizada por la mayoría de las CNNs de última generación para tareas de de audio mencionadas, como la detección de eventos acústicos y la clasificación de sonidos, por lo que deducimos que las tareas de estimación de los parámetros psicoacústicos mediante aprendizaje supervisado pueden considerarse más sencillas ya que partimos de secuencias de audio como entrada únicamente y no de vídeos por ejemplo que constan de imágenes de considerables dimensiones y sonido al mismo tiempo. Además el hecho de obtener una CNN de un tamaño considerablemente menor en lo que a número de filtros y dimensiones de los mismos representa, redunda en una velocidad de ejecución elevada, que es el fin buscado en esta Tesis Doctoral. Esto puede verse reflejado en los parámetros entrenables o *learnables* de la CNN diseñada que consta de 468555 parámetros entrenables frente a los 62.3 millones de parámetros entrenables que tiene el diseño de AlexNet. No obstante, como se demostrará en el siguiente apartado, ha sido el valor óptimo conseguido de parámetros entrenables, puesto que se ha probado a aumentar la complejidad de la CNN incluyendo más capas convolucionales sin lograr una mejora final del rendimiento.

La magnitud que se ha definido para ser minimizada durante el entrenamiento es el error cuadrático medio o RMSE (*Root Mean Square Error*) (3.3) que constituye por lo tanto nuestra función de coste.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{3.3}$$

Como optimizador hemos empleado el SGDM de *Stochastic Gradient Descent with Momentum*, con un momentum o impulso de de 0.9, que ayuda a acelerar la reducción de la función de coste acelerando así la convergencia al mínimo. Como el cálculo de derivadas parciales de la función de coste, que se ha visto en la Figura 3.3, respecto a cada uno de los pesos de la CNN y cada una de las observaciones, es inviable dado el número de pesos y observaciones, SGDM introduce un comportamiento estocástico limitando el cálculo a una sola derivada por observación (*batch*), optimizando así el proceso. De forma simplificada, el momentum o impulso se emplea acelerando el descenso en direcciones similares a las anteriores, pues el optimizador guarda un vector que representa la media de los anteriores vectores de descenso y si el nuevo vector es similar al vector de impulso, este amortigua la oscilación de la función de coste que se ralentiza, lo que acelera el descenso y por lo tanto la velocidad de convergencia. La tasa de aprendizaje inicial es de 10^{-3}, empleando un parámetro de regularización L2 de 10^{-4} que se aplica como parámetro a la función de coste, entrenando la red neuronal durante un máximo de 80 iteraciones (epochs) barajando los datos en cada iteración y con un tamaño de mini-bloque o mini-observación de 128 ejemplos. En la Figura 3.6 se puede ver cómo evoluciona el RMSE a la salida de la red neuronal mientras se reduce la función de coste (Loss) durante el proceso de entrenamiento. Este se ha realizado en el equipo informático descrito en la Tabla 2.2, como PC-1, equipado con la unidad de procesado de gráficos GPU o tarjeta gráfica NVIDIA GeForce GTX 1060 con 6 GB de memoria dedicada, que contribuye a acelerar el proceso de entrenamiento de la CNN. La mencionada GPU únicamente se ha empleado en el proceso de entrenamiento de la CNN y no ha participado en ninguna prueba ni con los algoritmos de cálculo directo no a la hora de ejecutar la CNN entrenada para predecir los parámetros psicoacústicos.

Como se ha mencionado anteriormente, la arquitectura propuesta finalmente responde a una serie de pruebas y correcciones precisas, realizadas sobre diferentes diseños y configuraciones de entrenamiento que han dado lugar al modelo óptimo presentado. Un ejemplo de este proceso se puede ver en la Tabla 3.3 donde se muestra la evolución del RMSE cometido

Tabla 3.2: Descripción de capas, CNN predicción parámetros psicoacústicos.

Capa	**Tamaño**	**Filtros**	**Paso**
Input	16000×1		
Convolutional S1	512×1	10	10
Batch Norm. S1			
ReLU. S1			
Max Pool S1	2×1		2
Convolutional S2	256×1	20	5
Batch Norm. S2			
ReLU. S2			
Max Pool S2	2×1		2
Convolutional S3	128×1	40	2
Batch Norm. S3			
ReLU. S3			
Max Pool S3	2×1		2
Convolutional S4	64×1	60	2
Batch Norm. S4			
ReLU. S4			
Max Pool S4	2×1		2
Convolutional S5	32×1	80	1
Batch Normalization S5			
ReLU			
Max Pool S5	2×1		2
Dropout 30 %			
Fully Connected	1×5		
Regression Output	1×5		

en la predicción del indicador general de molestia PA al emplear diferente número de etapas convolucionales en el diseño. Se observa cómo se alcanza el valor óptimo de RMSE con 5 etapas convolucionales, por lo que es el empleado en el diseño final.

3.2.3. Evaluación y resultados

El modelo diseñado de nuestra red neuronal convolucional que se ha descrito en el apartado anterior, una vez entrenado se puede exportar y ejecutar como si de cualquier algoritmo de cálculo se tratara, suministrándole secuencias de audio con las dimensiones definidas a su entrada y obteniendo la predicción de los parámetros psicoacústicos a la salida. Por lo

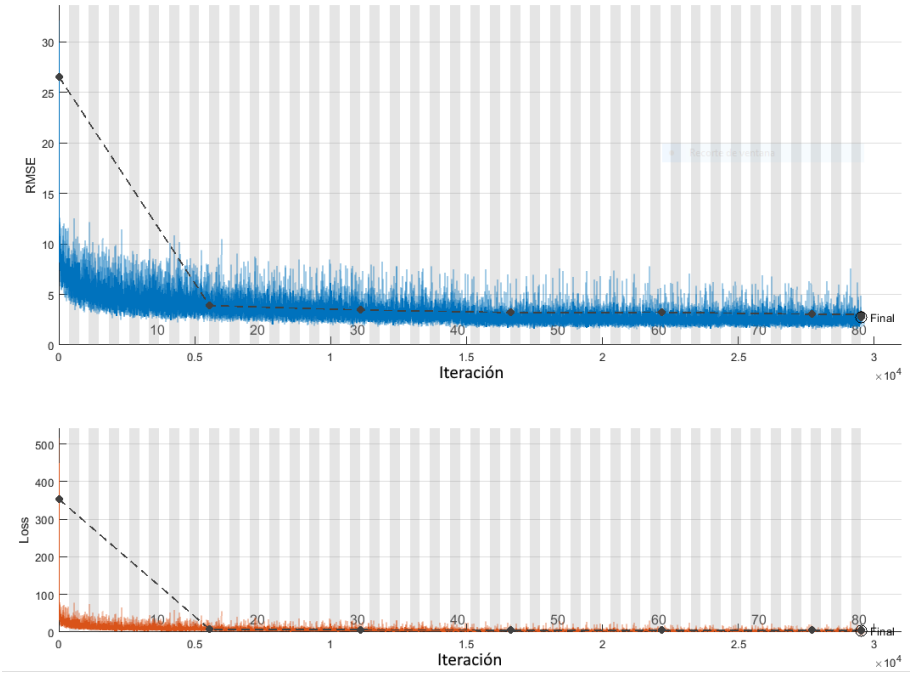


Figura 3.6: RMSE y Loss en proceso de entrenamiento.

Tabla 3.3: RMSE con diferente número de etapas convolucionales.

Nº Etapas Conv.	**RMSE predicción PA**
2 *etapas*	12.9810
3 *etapas*	4.5120
4 *etapas*	3.0390
5 *etapas*	2.5306
6 *etapas*	3.8480

tanto, hemos procedido a evaluar su rendimiento con una serie de pruebas que se detallan a continuación.

Análisis de pesos entrenados.

En primer lugar hemos analizado los pesos de la primera capa convolucional de la CNN, puesto que es donde se realiza la primera extracción de características del audio de entrada en bruto, tal y como se muestra en la Figura 3.7.

Observamos como los filtros aprendidos reflejan diferentes características espectro-temporales en forma de diferentes ondulaciones de frecuencia variable o modulaciones temporales. Por ejemplo, en el filtro F3 se observa claramente una modulación temporal de baja frecuencia sobre una frecuencia alta, mientras que los filtros sintonizados con frecuencias más bajas se pueden ver en F1, F4 y F5. Este comportamiento es similar algunas características relativas al cálculo de algunos parámetros psicoacústicos, como por ejemplo en el caso de *Loudness* o *Sharpness*, donde los niveles SPL se evalúan en función de las bandas de frecuencia según el patrón de percepción del oído humano, mientras que *Roughness* y *Fluctuation Strength* reflejan la presencia de modulaciones de baja frecuencia.

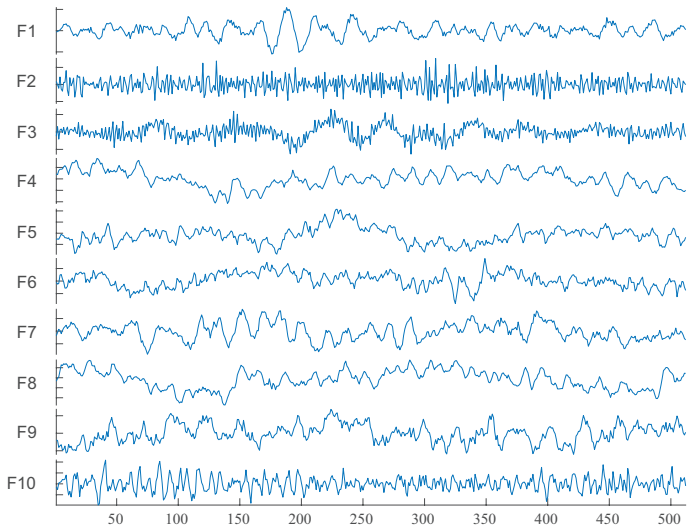


Figura 3.7: Pesos entrenados de la primera capa convolucional del modelo.

Evaluación de RMSE en predicción.

A continuación hemos evaluado la CNN con los tres conjuntos de datos de nuestra base de datos, Entrenamiento, Validación y Test, para predecir los 4 parámetros psicoacústicos (N, S, R y F) junto con el indicador general de molestia psicoacústica PA, y medir el RMSE comparando las predicciones y los parámetros calculados directamente de cada conjunto de datos. Los resultados del RMSE cometido se presentan en las Tablas 3.4, 3.5 y 3.6, junto con los valores máximos y mínimos de cada parámetro en cada conjunto de datos y el error porcentual con respecto al rango de los datos utilizados.

Tabla 3.4: RMSE de predicción, datos de Entrenamiento.

	N	**S**	**R**	**F**	**PA**
Valor Max.	67.27	3.73	25.06	13.37	99.89
Valor Min.	0.00	0.24	0.00	0.00	0.00
RMSE Pred.	0.8986	0.1998	0.2783	0.3224	1.5468
% Error	1.33 %	5.72 %	1.11 %	2.41 %	1.54 %

Tabla 3.5: RMSE de predicción, datos de Validación.

	N	**S**	**R**	**F**	**PA**
Valor Max.	78.15	3.37	26.97	12.49	97.40
Valor Min.	0.00	0.25	0.00	0.00	0.00
RMSE Pred.	0.9292	0.1981	0.4694	0.3622	2.5306
% Error	1.18 %	6.36 %	1.74 %	2.89 %	2.59 %

Observamos como, en casi todos los parámetros psicoacústicos, los mejores resultados se obtienen con los conjuntos de datos de entrenamiento y validación, como era de esperar, ya

Tabla 3.6: RMSE de predicción, datos de Test.

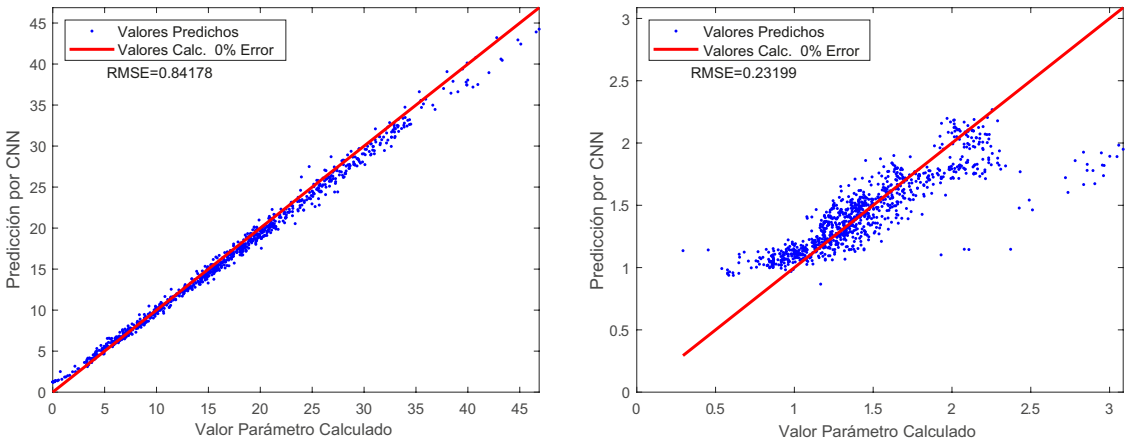
	N	**S**	**R**	**F**	**PA**
Valor Max.	46.87	3.08	12.81	6.48	98.62
Valor Min.	0.00	0.29	0.00	0.00	0.01
RMSE Pred.	0.8418	0.2320	0.2188	0.3329	1.4013
% Error	1.79 %	8.30 %	1.70 %	5.13 %	1.42 %

que han participado en el proceso de entrenamiento de la CNN. Presentan un RMSE muy por debajo del 3 % en la mayoría de casos, excediendo esta cifra únicamente para el caso de S, debido al desbalanceo de los datos de entrenamiento y validación. Además el indicador general de molestia PA, que se comporta como un parámetro psicoacústico resumen de todos los demás, arroja un RMSE máximo del 2.59 %, que es un resultado excelente. En la Tabla 3.6 se puede ver una buena descripción de la capacidad de la CNN para generalizar las predicciones a partir de datos desconocidos, pues muestra la evaluación empleando el conjunto de datos de test, que no han participado en el proceso de entrenamiento y validación de la red. De hecho, el error en predicción de N, s, R y F es mayor que a la hora de evaluar el resto de conjuntos de datos, pero no ocurre así con PA, que presenta un error ligeramente inferior a los otros casos, debido al proceso estocástico de elección de las señales de test, pero que demuestra las buenas capacidades de generalización de la CNN diseñada. Los valores mostrados son los óptimos alcanzados tras numerosas pruebas tanto con diferentes arquitecturas, como ya se ha mencionado, como con diferentes opciones de optimización y entrenamiento, teniendo en cuenta que la base de datos empleada aunque es muy realista, está considerablemente desbalanceada, como se ha visto en la Figura 3.4.

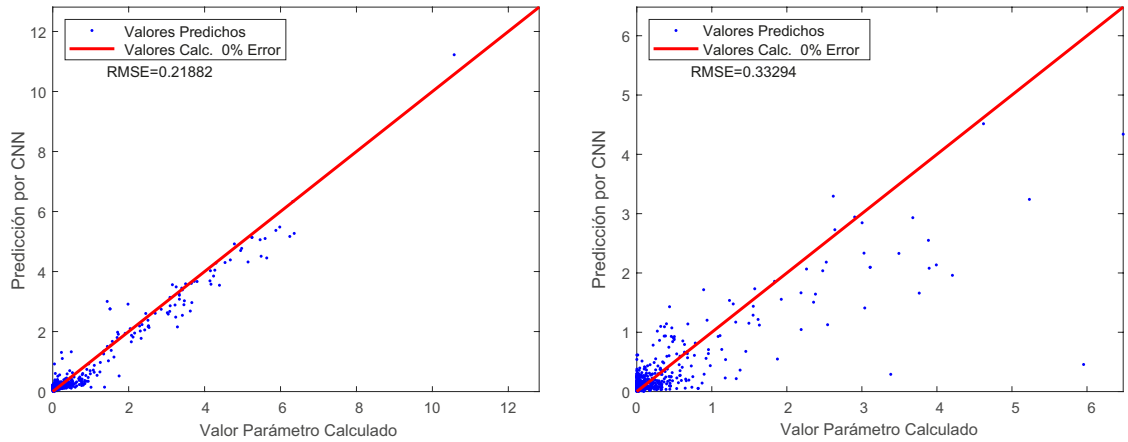
Distribución del error.

En las Figuras 3.8 y 3.9 se pueden ver los diagramas de dispersión con la predicción de los parámetros psicoacústicos N, S, R, F y PA (puntos azules) frente a los valores calculados de los mismos (lineas rojas), empleando los 1000 audios del conjunto de datos de test. Al comparar los resultados obtenidos para cada parámetro, se se observa que como se esperaba, el error de predicción tiende a ser mayor para aquellos parámetros que tienen un rango menos compensado de valores en el conjunto de datos de entrenamiento, como es el caso de S o F. No obstante, la CNN comete errores muy pequeños al predecir N o R. El error de predicción de PA también es muy bajo, demostrando nuestras pevisión, ya que el término que más influye en su cálculo es N, donde la CNN presenta muy buena precisión. Viendo la distribución de las predicciones dentro de cada parámetro, es interesante observar el comportamiento de PA en la Figura 3.9, que nos sirve como resumen de la predicción completa ya que está influenciado por el resto de parámetros psicoacústicos.

Sobre PA se ve claramente también cómo el error aumenta cuando presenta valores muy altos, pese a la buena precisión que demuestra. Este comportamiento es esperado, ya que en el conjunto de datos hay pocos ejemplos con valores de PA superiores a 50. En consecuencia, la predicción de PA tiende a dispersarse para valores superiores a ese rango. En general se puede ver este comportamiento reflejado en todos los parámetros teniendo en cuenta la distribución de datos de entrenamiento, lo que nos sirve de demostración empírica de cómo al proporcionar pocos datos en algún rango de valores la CNN no tiene casos suficientes para ajustar los coeficientes y aprender a generalizar por lo que comete más error en predicción.



(a) Loudness (b) Sharpness



(c) Roughness (d) Fluctuation Strength

Figura 3.8: Parámetros psicoacústicos. Predicción vs. valores calculados, conjunto de datos de test con RMSE final.

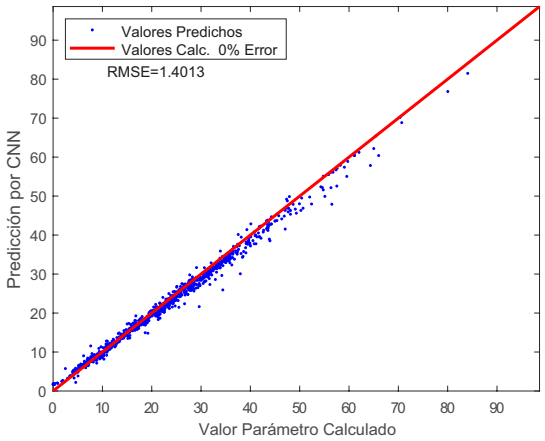


Figura 3.9: Predicción de PA vs. valores calculados, conjunto de datos de test con RMSE final.

Finalmente cabe denotar que pese a estas inusuales desviaciones, el modelo de CNN diseñado proporciona predicciones de los parámetros psicoacústicos junto con PA que recopila al resto de parámetros con una precisión excelente en la mayoría de casos.

Evaluación de RMSE en prueba de campo.

Hasta el momento se han descrito las pruebas realizadas empleando señales de audio de la base de datos definida, ya sea de los conjuntos de datos de Entrenamiento y Validación o de Test. Cabe recordar que la mayoría de estas secuencias de audio que componen nuestra base de datos han sido extraídas de la base de datos UrbanSound8k (59000 señales) mientras que una pequeña muestra pertenecen a grabaciones propias (1150 señales), por lo que se decidió realizar diversas grabaciones en la ciudad de Valencia para poder realizar una prueba de campo de la CNN con muchas más señales de audio de escenarios reales pertenecientes a escenarios diversos.

Tabla 3.7: RMSE de predicción, con datos independientes.

	N	**S**	**R**	**F**	**PA**
Valor Max.	72.32	2.94	6.03	5.58	85.4
Valor Min.	0.64	0.00	0.00	0.00	0.00
RMSE Pred.	0.93	0.2403	0.2609	0.2925	2.1704
% Error	1.29 %	10.43 %	4.32 %	5.23 %	2.54 %

El objetivo de esta prueba es validar la precisión de la CNN diseñada sobre los mencionados ejemplos de audio, que consisten en 2000 archivos de 1s de duración extraídos de grabaciones realizadas con una grabadora Zoom 4HN Pro. Estas grabaciones han sido realizadas en zonas de la ciudad y situaciones de actividad poco representadas en las secuencias de audio que han participado en la fase de entrenamiento de la CNN. Las señales de audio no han sido alteradas más allá de su remuestreo a 16 kHz y la calibración del micrófono, para no romper la filosofía de alimentar la CNN con audio sin procesar tal y como deseamos hacer en un nodo del sistema IoT.

En la Tabla 3.7 se muestran los resultados, que como se puede observar son similares a los obtenidos sobre el conjunto de datos de Validación, presentando ligeros incrementos de RMSE en los parámetros en los que ya se cometía más error. Esto evidencia la robustez de la CNN, que ha demostrado ser capaz de predecir con muy buena precisión los parámetros psicoacústicos de las nuevas grabaciones sonoras, a pesar de haber sido adquiridas en diferentes escenarios y con diferentes equipos de audio.

Tiempo de procesado.

Una vez verificada la precisión de la CNN diseñada y entrenada a la hora de predecir los parámetros de molestia psicoacústica, tanto en los conjuntos de datos que componen la base de datos creadas como en grabaciones externas a esta, las siguientes pruebas realizadas están relacionadas con el rendimiento de la CNN en ejecución. Es decir, ejecutándola en diferentes dispositivos para predecir los parámetros psicoacústicos a partir de diferentes señales de audio.

La primera evaluación que se ha llevado a cabo es del tiempo de predicción o ejecución. Ya que la motivación principal es reducir al máximo este tiempo para permitir el cálculo de los parámetros psicoacústicos en tiempo real y, a ser posible, en los mismos nodos del sistema IoT de monitorización. Recordemos que las señales de audio monitorizadas constan de 1 canal, con una duración de 1 segundo y una frecuencia de muestreo de 16 kHz. De ahí que las señales de la base de datos creada para el desarrollo de la red neuronal, tengan estas mismas características.

Los dispositivos empleados son los mismos que en el cálculo directo, descritos en la Tabla 2.2 de la sección 2.4, identificados como PC-1, PC-2 y Raspberry Pi 3B. Si bien en la mencionada sección 2.4.1 se detallaba el tiempo de procesado empleado en calcular cada parámetro psicoacústico, dado que la CNN proporciona en la capa de salida la predicción de todos los parámetros al mismo tiempo, la comparación se ha realizado teniendo en cuenta el tiempo total de cálculo directo frente al tiempo de predicción empleando la CNN. En la Tabla 3.8 se puede ver una comparativa entre el tiempo de cálculo directo frente al tiempo de predicción para 2 dipositivos, el ordenador de sobremesa PC-1 y la placa SBC Raspberry Pi 3B. Para realizar la prueba se han empleado las 1000 señales del conjunto de datos de test, descrito anteriormente. En este caso se ha empleado el algoritmo de cálculo directo con la primera optimización, que arroja un valor similar al mostrado anteriormente en la en la Tabla 2.4 de la sección 2.4.1.

Tabla 3.8: Parámetros psicoacústicos. Tiempos de calculo directo vs. predicción por CNN.

	μ [s]		σ^2 [s^2]	
	PC-1	**RPi3B**	**PC-1**	**RPi3B**
Cálculo directo, Opt. 1	1.3052	1.5182	$2{,}25 \cdot 10^{-3}$	$5{,}95 \cdot 10^{-3}$
Predicción CNN	0.0050	0.0259	$8{,}51 \cdot 10^{-6}$	$1{,}67 \cdot 10^{-4}$

Vemos de manera clara en la Tabla 3.8 que el tiempo medio μ empleado por la CNN para predecir el conjunto de parámetros psicoacústicos, es significativamente menor que el empleado por el cálculo directo. El tiempo medio empleado por la CNN es de 0.005 s, mientras que mediante cálculo directo tardamos 1.305 s, hablando en este caso del PC-1 de sobremesa. Esto representa un incremento de velocidad de más de 250 veces en este dispositivo. Además de los buenos resultados obtenidos para el ordenador PC-1, los obtenidos sobre la Raspberry Pi 3B, como sistema SBC son también excelentes. En la Raspberry Pi mientras que el cálculo directo emplea 1.518 s, la predicción realizada por la CNN, únicamente 0.0259 s, lo que representa casi 60 veces menos de tiempo y lo que es más importante, es inferior a 1 segundo (duración del audio de entrada) lo que permite el funcionamiento en tiempo real, que es el objetivo principal perseguido. En la Tabla 3.8, junto a los valores medios de tiempo μ obtenidos se muestra también la varianza media σ^2, para cálculo directo y predicción, que refleja la estabilidad del tiempo de cálculo requerido por la CNN para realizar la predicción de los parámetros a partir de las señales de audio sin procesar.

La varianza en el tiempo de cálculo requerido para el cálculo directo y para la predicción de la CNN se ve de forma más gráfica en la Figura 3.10, donde se muestran para cada señal de audio analizada. Aquí, se puede observar que los tiempos requeridos para la predicción por la CNN son prácticamente constantes, mientras que los tiempos obtenidos por cálculo directo muestran más dependencia de la complejidad del sonido de entrada.

En la Tabla 3.9 podemos ver la comparativa de tiempos medios μ de cálculo directo y predicción por la CNN en segundos, en este caso empleando la ultima optimización de código para el cálculo directo. Se aprecia que aunque la optimización consigue mejorar los tiempos de cálculo directo, como ya se describió anteriormente, la predicción de la CNN es hasta 150 veces más rápida en el caso del equipo que presenta el mejor rendimiento, el PC-1. En el caso de dispositivos donde el procesador tiene recursos más modestos, como en el PC-2 o la placa SBC Raspberry Pi, la diferencia entre las optimizaciones del cálculo directo es mucho más limitada y la diferencia con el tiempo de predicción de la CNN se mantiene en una clara ventaja para esta última. En la Raspberry Pi, el mejor tiempo obtenido para el cálculo

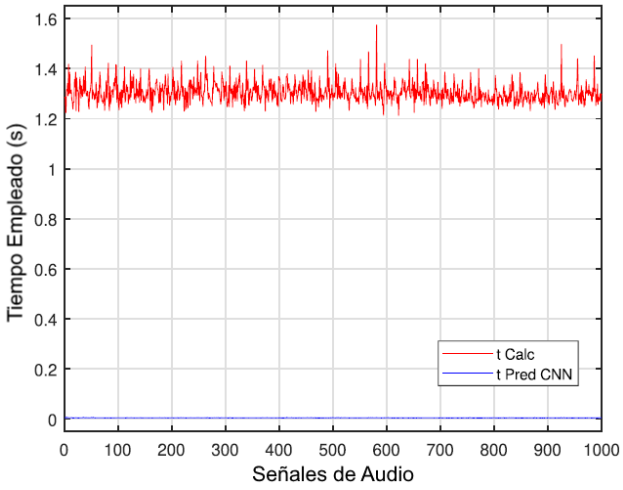


Figura 3.10: Parámetros psicoacústicos. Tiempos calculo directo vs. predicción por CNN en PC-1.

Tabla 3.9: Parámetros psicoacústicos. Tiempos de calculo directo (Opt.2) vs. predicción por CNN.

	μ [s]		
	PC-1	**PC-2**	**RPi3B**
Cálculo directo, Opt. 2	0.6701	1.2163	1.4839
Predicción CNN	0.0048	0.0181	0.0258

directo de los parámetros psicoacústicos es de 1.48 s para cada señal de audio de entrada, mientas que la predicción de la CNN se realiza en unos 0.026 s, por lo que sigue siendo casi 60 veces más rápida. Una vez optimizadas las operaciones de lectura de datos del disco duro y carga de elementos en memoria. prácticamente el total de operaciones que quedan se realizan en el procesador o CPU del equipo, por lo que las diferencias las marca este elemento. Por lo tanto, los dispositivos con una CPU de prestaciones más modestas respecto a memorias caché o frecuencia de funcionamiento y no tanto a número de núcleos, se ven más afectadas al realizar el cálculo de los parámetros psicoacústicos. En estos sistemas las redes neuronales convolucionales marcan la diferencia tal y como se ha demostrado en esta investigación. Cabe mencionar que en ningún caso se ha empleado ninguna GPU o tarjeta gráfica (dedicada o integrada en la CPU) para realizar ningún cálculo en ningún dispositivo, realizándose los mismos en la CPU únicamente.

Como se ha descrito, el excelente resultado habilita automáticamente al sistema IoT diseñado para trabajar en un modo de monitorización continua en tiempo real sin interrupciones para calcular o despreciar tiempo sin monitorizar. En la práctica esto representa más ventajas todavía. El primer elemento que se ha reducido es el buffer de memoria que ha pasado a ser de apenas 4.2 kB , ya que no es necesario almacenar grandes secuencias de audio de decenas de MB a la espera de ser analizadas ni tampoco es necesaria una gran unidad de almacenamiento local. En la Tabla 3.10 re puede ver una comparativa del uso de RAM por el algoritmo de cálculo directo frente a la RAM usada por la CNN. Como puede verse, pese a no ser un elemento vital, se reduce en gran medida la cantidad de RAM empleada al realizar las predicciones con la CNN diseñada, lo que redunda en beneficio general del sistema, sobretodo en casos críticos donde haya que mantener datos de gran volumen o aplicaciones en memoria.

Tabla 3.10: Parámetros psicoacústicos. Uso de RAM, cálculo directo vs. predicción CNN.

	PC-1 (MB)	**RPi3B** (MB)
Cálculo directo, Opt. 2	83.81	83.34
Predicción CNN	12.57	18.36

Otro beneficio inherente a realizar la predicción de los parámetros psicoacústicos en el mismo nodo es que se elimina por completo la necesidad de mandar secuencias de audio para realizar los cálculos en el equipo central o nodo de control del sistema IoT. Se evitan así posibles problemas con la Ley Orgánica de Protección de Datos de Carácter Personal, pues lo único que se transmite y almacena son los parámetros psicoacústicos (incluido el indicador general de molestia PA), que en ningún caso permiten la reconstrucción de las señales originales de audio, que pueden ser eliminadas de memoria en cuanto se calculan los mismos.

Una consecuencia beneficiosa más de esta reducción del tiempo de procesado, es la reducción de consumo. Al realizarse la predicción por la CNN en unos 26 ms en el caso de la Rasperry Pi, antes de procesar la siguiente secuencia de audio, únicamente hay que enviar el valor de los parámetros por red al nodo de control y se puede establecer un estado de reposo o bajo consumo (a nivel de procesador) hasta que se deba procesar la siguiente secuencia de audio, pues el consumo de CPU para monitorizar el audio del micrófono es muy reducido. La reducción de consumo resulta directamente en una mayor autonomía de las baterías, y por lo tanto incrementa el número de horas que el sistema IoT puede estar monitorizando de forma continua sin conexión a la red eléctrica.

Estos datos se han publicado en los artículos de revista *Computation of Psycho-Acoustic Annoyance Using Deep Neural Networks* [24] y *Enabling Real-Time Computation of Psycho-Acoustic Parameters in Acoustic Sensors Using Convolutional Neural Networks* [25], en distintas fases de evolución, incluidos en la compilación de publicaciones de esta Tesis Doctoral como Anexos A y B.

En definitiva, aplicando una red neuronal convolucional entrenada mediante aprendizaje profundo hemos conseguido predecir los parámetros psicoacústicos hasta 150 veces más rápido que empleando el cálculo directo en un equipo informático comercial y hasta 60 veces más rápido en una placa SBC como es la Raspberry Pi 3B. Al mismo tiempo se mantiene una tasa de error muy baja en las predicciones como se ha visto anteriormente. El buen desempeño de la CNN diseñada permite realizar las predicciones de los parámetros psicoacústicos dento del mismo nodo del sistema IoT, de manera que se descarga el nodo central de las tareas de cálculo y la red de tráfico de datos excesivo, al mismo tiempo que se reduce el consumo y se aumenta la autonomía. El sistema IoT diseñado puede por lo tanto monitorizar de forma continua sin interrupciones ni problemas de desbordamiento de memoria. Estos resultados en conjunto permiten cumplir con los objetivos de esta Tesis Doctoral, al menos en lo que respecta a los parámetros psicoacústicos. Por lo tanto el siguiente paso natural ha sido aplicar el mismo modus operandi a los parámetros acústicos de sala, como se describe en la siguiente sección.

3.3. CNN aplicada a parámetros acústicos de sala

Tras los buenos resultados obtenidos aplicando las redes neuronales convolucionales a la predicción de los parámetros psicoacústicos, se ha planteado la opción de emplearlas también para simplificar el proceso de cálculo de los parámetros acústicos de sala, que comprenden el segundo conjunto de parámetros acústicos seleccionados. Así pues, en esta sección se describe la arquitectura de la CNN diseñada para predecir los parámetros de sala, la base de datos creada a medida para tal fin, la configuración de entrenamiento y finalmente las pruebas realizadas al modelo propuesto así como los resultados obtenidos.

Recordemos que el conjunto de parámetros de sala seleccionado comprende $RT60$, $C50$, $C80$, STI y SII. Los 4 primeros se calculan a partir de la respuesta impulsiva de la sala, y el quinto a partir de una señal de habla reproducida en la sala a estudiar. Por lo tanto es necesario obtener la respuesta impulsiva de la sala mediante el empleo de fuentes de sonido y la reproducción de barridos en frecuencia o señales MLS, lo que dificulta y ralentiza mucho el análisis rápido de los parámetros de sala. Uno de los objetivos principales de esta Tesis es el desarrollo de nuevos métodos de monitorización y análisis acústico, incorporando por ejemplo las redes de sensores inalámbricos, tal y como se he hecho con el sistema IoT diseñado. Si bien, es verdad que facilita en gran medida la instalación del sistema (frente a uno cableado), la adquisición de señales y el cálculo de los parámetros de sala, no nos libera de la tarea de calcular la respuesta impulsiva.

Como ya se ha mencionando, la velocidad de cálculo no es esencial en este caso, pues los valores de los parámetros acústicos de sala no cambian rápidamente con el tiempo, sin embargo la necesidad de obtener la respuesta impulsiva emplea casi el 75 % del tiempo de cálculo en el mejor de los casos. Cabe recordar que para establecer las mismas condiciones para todos los dispositivos, se han empleado señales MLS de 20 s de duración, cumpliendo criterios de la norma ISO-3382 para obtener estas respuestas impulsivas, sin embargo hemos empleado señales mucho más largas en diferentes pruebas que se han realizado a lo largo de esta investigación, por lo que este porcentaje ha demostrado ser mucho más elevado en esos casos. Debido a ello, optimizar este elemento ha centrado los esfuerzos de esta parte de la investigación, alentada además por los buenos resultados obtenidos al aplicar las redes neuronales convolucionales a la predicción de los parámetros psicoacústicos.

3.3.1. Base de datos

Como en el acaso de los parámetros psicoacústicos y de forma general para cualquier desarrollo de un modelo de CNN basado en aprendizaje profundo es necesario disponer de datos suficientes tanto en cantidad como en características para poder llevar a cabo el entrenamiento. El caso del modelo diseñado para predecir los parámetros acústicos de sala es más complejo que el dedicado a los parámetros psicoacusticos, pues si bien estos se podían calcular todos a partir del una señal de audio grabada en la zona a monitorizar, para los parámetros de sala se precisa de la respuesta impulsiva de la sala para 4 parámetos y de una señal de habla para el último parámetros (SII). En nuestro caso deseamos que la CNN sea capaza de predecir los parámetros acústicos de sala a partir de una señal de habla grabada en el recinto a analizar, por lo que necesitamos disponer de numerosas señales de habla grabadas en un gran número de salas para poder disponer de un conjunto suficientemente amplio de señales para entrenar, junto con el valor de los parámetros acústicos de sala para cada caso.

Dado que se desea una gran capacidad de generalización, los datos de entrenamiento deben contener grabaciones correspondientes a salas de características muy diferentes. Se debe tener en cuenta que encontrar un número elevado de salas con las características geométri-

cas y constructivas adecuadas para generar una diversidad de señales suficiente es una tarea complicada. Por ello, para crear nuestra base de datos se ha optado por emplear un conjunto de respuestas impulsivas de diferentes salas para convolucionar con otro conjunto de señales de habla grabadas de forma anecoica.

En el caso de las respuestas impulsivas, se ha generado un primer subconjunto correspondiente a 15 salas distintas de forma sintética, empleando el método de las imágenes [96, 97]. Estas respuestas corresponden a 15 salas de geometría rectangular de tamaño creciente, desde 1 m^2 a 1600 m^2, donde se han ajustado tanto los coeficientes de reflexión como la geometría para obtener diferentes valores representativos de RT60, de 0.1 a 1.5 segundos. Se ha simulado emplazando la fuente de sonido y los micrófonos receptores en 20 posiciones distintas por sala para disponer de diversas distancias de fuente - micrófono y obtener un total de 300 respuestas, que ayudan a tener más rango de valores en el caso de parámetros directamente relacionados con la distancia del oyente a la fuente de sonido, como es SII. En la Figura 3.11 se puede ver un ejemplo de situación de fuentes de sonido (marcados como puntos azules) y micrófonos (marcados con círculos rojos) en salas de diferentes dimensiones. A la hora de elegir las posiciones de estos, se ha seguido como norma general las indicaciones de la norma ISO-3382 para la obtención de respuestas impulsivas reales, orientando la ubicación de elementos a un esquema de sala tipo *escenario / público*, con una zona destinada a la presentación de contenido y otra al público asistente.

Este primer subconjunto se ha completado con 10 respuestas impulsivas de 10 salas reales extraídas del repositorio OpenAir (*Open Acoustic Impulse Response Library*) [1] y que se han seleccionado de todas las disponibles las correspondinetes a espacios cerrados, en este caso de diversas geometrías irregulares. De esta manera se ha formado un conjunto de 310 respuestas impulsivas en total, correspondientes a 25 salas de naturaleza diversa, para confeccionar la base de datos empleada.

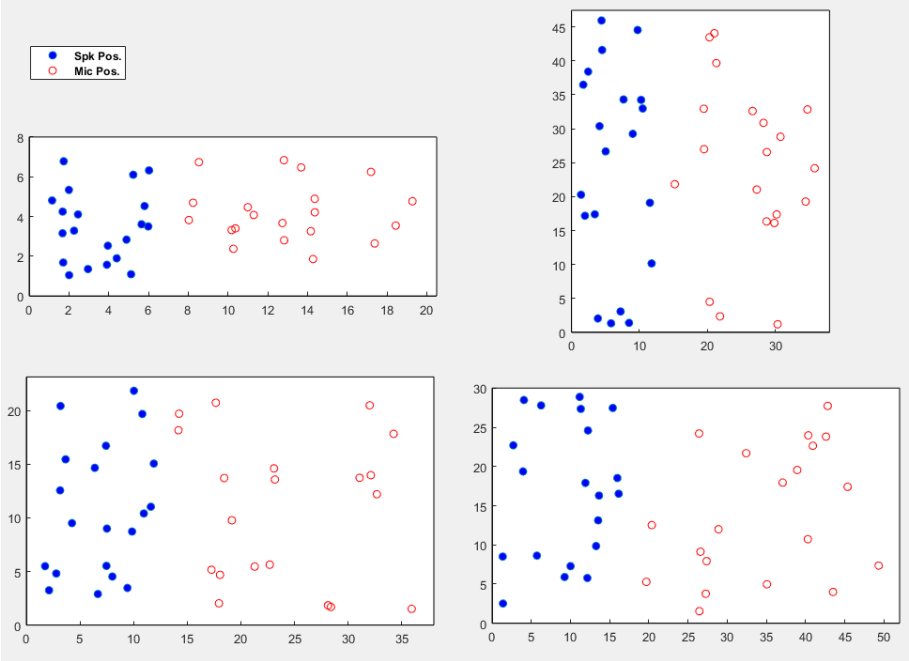


Figura 3.11: Ejemplos de posicionamiento de fuentes (puntos azules) y micrófonos(círculos rojos) para la generación respuestas impulsivas sintéticas.

El siguiente elemento para confeccionar la base de datos son las señales de voz. Estas

[1] https://www.openairlib.net/

han sido extraídas de la base de datos *DARPA-TIMIT acoustic-phonetic continuous speech database* [85], que contiene 6.300 grabaciones de personas que pronuncian diferentes frases en 8 dialectos principales del idioma inglés americano. Para conformar la base de datos se ha realizado un pre-procesado en 3 pasos:

1. Remuestreo de las señales de voz a 16 kHz para tener frecuencias de muestreo uniformes en todo el conjunto de datos.

2. Convolución de las señales de habla con las respuestas impulsivas.

3. Recorte o extensión de las señales obtenidas a para fijar una duración uniforme de 3.5 segundos para todas, lo que define señales de 56000 muestras.

A la hora de formar definitivamente la base de datos, empleando las respuestas impulsivas y las señales de voz seleccionadas, hemos calculado los parámetros de sala pertinentes de las respuestas impulsivas antes de la convolución ($RT60$, $C50$, $C80$ y STI) y el parámetro SII de la señal de voz obtenida después de convolucionar. Seguidamente hemos almacenando las señales de audio obtenidas y los datos identificativos junto con los parámetros correspondientes a cada una. De esta manera no es necesario realizar el pre-procesado ni el cálculo de los parámetros en el momento de entrenar la CNN.

Esto ha permitido obtener un total de 69305 señales, sin embargo diversas pruebas demostraron que emplear todas en el entrenamiento de la CNN producía sobre-ajuste u *overfitting*, por lo que se ha fijado el número óptimo de señales a emplear en 30000. Estas 30000 señales son las únicas que participan en el proceso de entrenamiento y se han dividido en 2 conjuntos, el 80 % para entrenamiento (24.000 señales) y el 20 % para validación (6.000 señales). Para poder realizar una prueba más exhaustiva e independiente de la CNN entrenada, hemos definido un conjunto más con 1000 señales de test. Este conjunto de señales no participa en el proceso de entrenamiento y validación de la red neuronal y para garantizar su independencia hemos empleado 70 señales de voz más que no están presentes en los otros conjuntos y 15 respuestas impulsivas nuevas (10 de OpenAir y 5 sintéticas), lo que nos ha permitido generar 1050 señales de donde hemos seleccionado aleatoriamente 1000.

Por lo tanto nuestra base de datos está formada por 3 conjuntos de señales de habla modificadas por las características de diversas salas con la distribución siguiente:

1. **Entrenamiento**, con 24000 señales.

2. **Validación**, con 6000 señales.

3. **Test**, con 1000 señales.

Sobre la definición del número de señales de voz y de respuestas impulsivas a emplear para conformar la base de datos, dado que 25 respuestas impulsivas puede parecer un número reducido, cabe destacar que es resultado de las pruebas preliminares realizadas en el proceso de diseño. Durante estas se pudo observar que la precisión final obtenida sobre los datos de validación dependía mucho más de la variabilidad y diversidad del contenido del habla en las señales de entrenamiento que de la variabilidad de las respuestas impulsivas con parámetros de sala similares. Por ello se ha definido el conjunto de datos de entrenamiento para conseguir más variabilidad en las señales de habla que en el número de respuestas impulsivas empleadas para simular diferentes entornos acústicos. En estas pruebas preliminares se ha tenido en cuenta también que el modelo podría sobre-ajustarse a los datos generados a partir de respuestas de impulso sintéticas, por lo que los datos de entrenamiento se equilibraron incluyendo un número similar de respuestas de impulso reales, tal y como se ha descrito. Mediante estos test se ha establecido también el número óptimo de señales a emplear en el entrenamiento y validación que ayudan a evitar este sobre-ajuste.

Cabe mencionar que de todos los trabajos encontrados que se centran en la predicción de parámetros acústicos de sala empleando modelos de CNN [90, 98, 99, 100, 101] y que se describirán más adelante para comparar rendimientos, únicamente el último predice más de 2 parámetros acústicos y emplea un total de 29000 señales para realizar el entrenamiento de la CNN. Es curioso que esta cifra está cerca de la que hemos seleccionado en nuestro caso como óptima, teniendo en cuenta que no nos hemos basado en esa investigación pues el mencionado estudio se ha publicado recientemente cuando estábamos realizando ya las pruebas de campo de nuestro modelo de CNN.

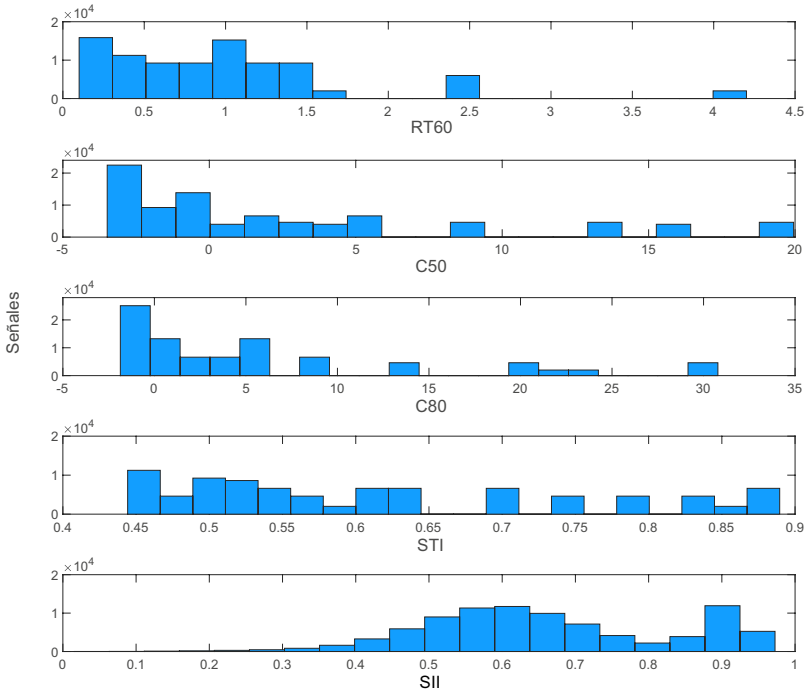


Figura 3.12: Histograma de los valores calculados de $RT60$, $C50$, $C80$, STI y SII de la base de datos creada.

En la Figura 3.12 se pueden ver los histogramas de los parámetros de sala calculados para las señales de voz incluidas en la base de datos. Observamos por ejemplo que en el caso de $RT60$ es donde se produce más desequilibrio en el número de señales, pues hay relativamente pocas para algunos valores de reverberación (3 segundos por ejemplo). En general los primeros 4 parámetros, como se ha mencionado, se calculan a partir de las respuestas impulsivas por lo que presentan una distribución discreta en sus valores, mientras que SII, calculado a partir de la señal de voz final presenta una distribución continua con valores presentes en casi todo el rango, echándose de menos únicamente valores por debajo de 0.2, debido a que obtener valores tan reducidos de inteligibilidad es complicado en salas con dimensiones y distancias reales de oyente a orador.

Como en el caso de la base de datos empleada para entrenar la CNN de los parámetros psicoacústicos, no hemos empleado técnicas de aumento de datos para incrementar el número de señales. Como se ha mencionado, en las pruebas preliminares ya se observó que este aumento producía rápidamente un sobre-ajuste de la respuesta de la CNN y además no permitía obtener señales con parámetros ubicados en las zonas de interés para mejorar la distribución de los conjuntos, como por ejemplo $RT60$ de 3 segundos que presenta pocas señales, sino que contribuía a aumentar el número de señales en los rangos donde ya están presentes.

3.3.2. Diseño, configuración y entrenamiento

El diseño de la CNN orientada a predecir los parámetros acústicos de sala se basa en el descrito anteriormente enfocado a parámetros relativos a molestia psicoacústica. Por lo tanto se ha repetido el esquema de emplear filtros cuyo tamaño disminuye desde las capas de entrada hasta la salida, mientras que su número aumenta. La selección de hiperparámetros que estructuran las capas convolucionales se ha realizado probando numerosas arquitecturas hasta lograr la máxima precisión alcanzada en la respuesta. El esquema básico que ha presentado la mejor respuesta se basa en unidades convolucionales formadas por 3 capas: Convolucional + Capa ReLU + Capa MaxPool. La capa de convolución temporal es seguida por una capa de activación *Rectified Linear Unit* (ReLU) que pone a 0 cualquier entrada negativa. La siguiente capa (MaxPool) realiza un diezmado (downsampling) transmitiendo únicamente el máximo de los elementos indicados. Esta unidad convolucional se repite 4 veces y forma el núcleo de la red neuronal diseñada, que se puede ver esquemáticamente en la Figura 3.13

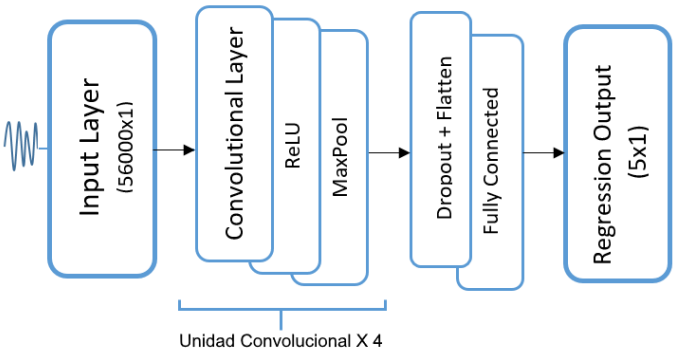


Figura 3.13: Diseño CNN parámetros acústicos de sala.

Partiendo de una señal de voz se llega al conjunto descrito de 4 unidades convolucionales. Les sigue una capa DropOut con una probabilidad de abandono de 0.3 que evita el sobreajuste y una capa Flatten o de aplanamiento que convierte las matrices en un vector. Esta capa conduce a la capa FullConnected que reduce la salida a 5 elementos seguida de la capa final de activación por regresión lineal que proporciona la predicción de $RT60$, $C50$, $C80$, STI y SII.

La descripción detallada de los parámetros de cada capa se puede ver en la Tabla 3.11. Desde la capa de entrada donde llegan las 56000 muestras de la señal de voz se puede ver cómo a lo largo de la CNN el número de filtros convolucionales aumenta al mismo tiempo que disminuye su tamaño para converger en la capa de salida con las 5 predicciones de los parámetros de sala.

La profundidad de la red ha sido diseñada hasta alcanzar la mejor precisión, siendo de destacar que es mucho menor que la utilizada por la mayoría de las CNNs de última generación aplicadas a acústica en tareas como la clasificación de sonidos o la detección de eventos, sin embargo es considerablemente superior a los únicos que hemos encontrado enfocados a la predicción de parámetros de sala de forma individual o en conjunto. Nuestro modelo de CNN consta de 322955 parámetros entrenables, que si bien es muy reducido comparado con AlexNet (62.3 millones de parámetros entrenables), es 100 veces superior a otros modelos de CNN más cercanos al nuestro, como el presentado en el estudio [100], que declara 8737 parámetros entrenables, si bien para predecir 2 parámetros acústicos únicamente ($RT60$ y DRR, *Direct to Reverberant Ratio*). Como se verá más adelante el hecho de tener una CNN de tamaño comedido redundará en una velocidad de ejecución elevada que es una característica

muy interesante nuestro caso.

Tabla 3.11: Descripción de capas, CNN predicción parámetros acústicos de sala.

Capa	**Tamaño**	**Filtros**	**Paso**
Input	56000×1		
Convolutional S1	512×1	10	10
ReLU. S1			
Max Pool S1	2×1		2
Convolutional S2	256×1	20	5
ReLU. S2			
Max Pool S2	2×1		2
Convolutional S3	128×1	40	2
ReLU. S3			
Max Pool S3	2×1		2
Convolutional S4	64×1	60	1
ReLU. S4			
Max Pool S4	2×1		2
Dropout 30 %			
Flatten			
Fully Connected	1×5		
Regression Output	1×5		

La función de coste que hemos definido es el error cuadrático medio o MSE (*Root Mean Square Error*) (3.4) que será minimizada durante el proceso de entrenamiento. El MSE es muy sensible a los valores atípicos que difieren de la media, por lo tanto se ajusta bien a los problemas basados en regresión, donde se espera que la distribución de la salida condicionada a los datos de entrada sea gaussiana, y donde pretendemos penalizar más (cuadráticamente) los valores de error más grandes que los pequeños. Esto nos permite minimizar los errores más grandes con mayor prioridad sobre los más pequeños durante el proceso de entrenamiento. Como resultado, la CNN se adapta al uso final en el que grandes errores en la predicción de los parámetros acústicos llevarían a un análisis erróneo del comportamiento de la sala analizada.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{3.4}$$

Como optimizador, de todos los ensayados, el que mejor resultado ha devuelto ha sido el Adam (derivado de *adaptive moment estimation*), que se suele emplear mucho en aplicaciones de deep learning enfocadas al procesado de lenguaje natural donde ha demostrado un excelente desempeño. Este optimizador es una extensión del SGD descrito anteriormente, donde en lugar de emplear una tasa de aprendizaje fija para todas las actualizaciones de pesos,

que no cambia durante el entrenamiento, el optimizador Adam mantiene una tasa para cada parámetro entrenable y adapta esta tasa al desarrollo del proceso de aprendizaje. Para ello se basa en el primer momento promedio (la media),como otros optimizadores, pero haciendo uso además de la media de los segundos momentos de los gradientes (o varianza descentrada), modificados por los parámetros β_1 y β_2 respectivamente. Para la configuración del optimizador Adam hemos definido los hiperparámetros siguientes: tasa de aprendizaje (α) de 0.001, β_1 de 0.9 y β_2 de 0.999.

Otros hiperparámetros que se han definido son la tasa de regularización L2 que se ha establecido en 0.001, el mini-bloque, establecido en 512 con barajado de datos en cada iteración y número máximo de iteraciones o epochs establecido en 350, empleando una función que almacena el modelo cada vez que se reduce el MSE en validación durante el proceso de entrenamiento.

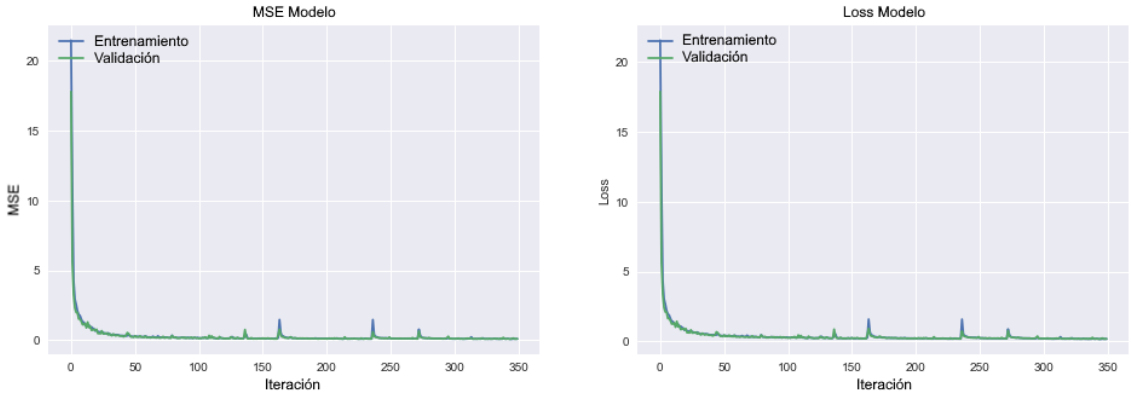


Figura 3.14: MSE y Función de Coste (Loss) en proceso de entrenamiento.

En la Figura 3.14 se puede ver cómo evoluciona el MSE a la salida de la red neuronal mientras se reduce la función de coste (Loss) durante el proceso de entrenamiento. El entrenamiento del model de CNN se ha realizado en el equipo informático descrito en la Tabla 2.2, como PC-1, como en el caso anterior y que está equipado con una tarjeta gráfica NVIDIA GeForce GTX 1060 con 6 GB de memoria dedicada. Cabe mencionar que la GPU únicamente se ha empleado en el proceso de entrenamiento de la CNN y no para ninguna otra prueba realizada sobre la CNN entrenada.

Como en el caso anterior con el modelo de CNN dedicado a parámetros psicoacústicos, se han realizado numerosas pruebas variando los hiperparámetros para el entrenamiento hasta encontrar el ajuste perfecto y obtener el mínimo MSE a la salida de la red neuronal. Además se han ensayado diferentes estructuras, número y dimensiones de capas. Desde el primer resultado que arrojaba un 16.2946 de MSE medio, hemos realizado los ajustes necesarios hasta llegar a un MSE 0.0856 de media para la predicción de los 5 parámetros sobre el conjunto de datos de validación.

3.3.3. Evaluación y resultados

Una vez concluido el proceso de entrenamiento del modelo de CNN descrito, se ha procedido a evaluar su rendimiento de forma exhaustiva, tanto en términos de precisión como de velocidad de ejecución sobre diferentes dispositivos y escenarios, como se describe a continuación.

Análisis de pesos entrenados.

En la Figura 3.15 se puede ver una representación de los parámetros o pesos entrenados correspondientes a la primera capa convolucional de la CNN. Puesto que es un modelo basado en deep learning al que se le suministra una señal de audio en bruto a la entrada, es en estas primeras capas donde la CNN implementa por sí sola la primera extracción de características.

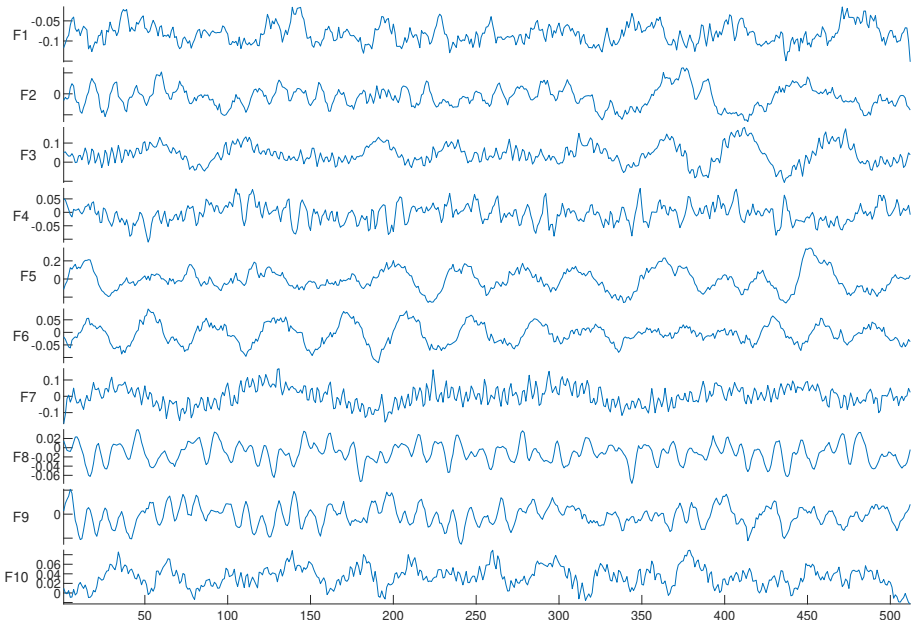


Figura 3.15: Pesos entrenados de la primera capa convolucional del modelo.

Se observa como los filtros aprendidos presentan diferentes características espectro-temporales que se aprecian en la forma de onda que toman los pesos aprendidos. En los filtros F3, F7 o F10 se pueden ver claramente modulaciones en frecuencia con diferentes frecuencias moduladoras, situándose la más baja en F7 por ejemplo. Al estar viendo los filtros en el dominio temporal, podemos apreciar también cómo se realiza un filtrado por secciones de la señal de entrada, como por ejemplo en el filtro F2, donde hasta la muestra 350 se aplica un filtrado con unas características y de ahí al coeficiente 512, con otras, realizándose así un pseudo-análisis por tramos de la sección de entrada, derivado seguramente por la presencia de señales de voz seguidas de algún tramo de silencio en muchas de las señales presentes en la base de datos. Observamos una analogía entre este comportamiento que implementa de manera automática la CNN durante el proceso de entrenamiento y la extracción de características espectro-temporales que realizaríamos nosotros en una análisis de la señal y que también implementan los algoritmos de cálculo de los parámetros acústicos de sala.

Evaluación de precisión en predicción.

La siguiente serie de pruebas están relacionadas con la evaluación del error cometido en predicción por la CNN entrenada. Midiendo este sobre el conjunto de datos de validación y test al realizar las predicciones de los 5 parámetros acústicos de sala.

En este caso, hemos evaluado los errores únicamente sobre los conjuntos de datos de validación y test sin evaluar sobre el conjunto de entrenamiento, pues con ellos tendremos una muestra de un conjunto de datos que ha participado en el proceso de entrenamiento y de un conjunto de datos totalmente independiente. Para apreciar mejor la precisión de la CNN se han evaluado diferentes expresiones del error experimental cometido en predicción. El primero evaluado ha sido el error cuadrático medio (MSE), empleado como función de

coste durante el entrenamiento según la expresión 3.4 de la sección 3.3.2). Se ha evaluado también el error medio absoluto o MAE (*Mean Absolute Error*) calculado según 3.5 que nos es útil para cuantificar la precisión de la CNN observando el error cometido de forma absoluta, y que representaría la distancia vertical promedio de las predicciones a la recta ideal de los valores correctos calculados. El siguiente error calculado es el error relativo medio (3.6) que cuantifica el porcentaje de error cometido frente a los valores reales para cada parámetro. Para apreciar mejor los resultados cabe recordar que el MSE se mide en las mismas unidades al cuadrado que el parámetro que se estima, por ejemplo para $RT60$ serían $[s^2]$. El MAE se mide en las mismas unidades que el parámetro estimado y el MRE en nuestro caso lo expresamos en porcentaje al ser una relación y carecer de unidades.

$$MAE = \frac{\sum_{i=1}^{n} |\hat{Y}_i - Y_i|}{n} \tag{3.5}$$

$$MRE(\,\%) = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{Y}_i - Y_i|}{\hat{Y}_i} \cdot 100 \tag{3.6}$$

EN las Tablas 3.12 y 3.13, se pueen ver los valores descritos de MSE, MAE y MRE en porcentaje por parámetro acústico, para los datos de validación y test respectivamente. Junto a ellos se muestra también la varianza media del error (σ) y el valor de la correlación cruzada o correlación Pearson media (ρ) entre las predicciones y los valores correctos.

Tabla 3.12: Evaluación de error en predicción, datos de Validación.

	RT60 (s)	**C50 (dB)**	**C80 (dB)**	**STI**	**SII**	**Media**
MSE	0.0117	0.1981	0.2181	0.0002	0.0005	0.0857
MAE	0.0569	0.2630	0.2850	0.0088	0.0128	0.1253
MRE (%)	4.88	7.37	9.06	1.46	1.42	4.84
Error σ	0.0935	0.2970	0.3020	0.0097	0.0189	0.1442
ρ	0.9959	0.9970	0.9979	0.9937	0.9861	0.9941

Tabla 3.13: Evaluación de error en predicción, datos de Test.

	RT60 (s)	**C50 (dB)**	**C80 (dB)**	**STI**	**SII**	**Media**
MSE	0.0134	0.2100	0.2235	0.0003	0.0008	0.0896
MAE	0.0652	0.2788	0.2921	0.0132	0.0205	0.1339
MRE (%)	5.59	7.82	9.29	2.19	2.28	5.43
Error σ	0.0957	0.3636	0.3718	0.0135	0.0211	0.1732
ρ	0.9951	0.9969	0.9978	0.9919	0.9697	0.9903

Como es lógico, los errores cometidos en predicción sobre el conjunto de señales de validación son inferiores, debido a que se emplean para validar el entrenamiento de la red neuronal y por lo tanto participan de este proceso activamente, pero nos sirve para comprobar que el comportamiento es correcto y lógico. Los datos de test no han participado de ninguna forma en el proceso de entrenamiento, por lo que son una muestra mucho más precisa y objetiva

del desempeño de la CNN a la hora de predecir los parámetros acústicos de sala.

Como se puede ver en la Tabla 3.13, observando el MRE en porcentaje, obtenemos un 5.43 % de error de media para los 5 parámetros, pero de forma individual en ningún caso se supera el 10 % de error para ningún parámetro, acercándose únicamente a esta cifra las predicciones de $C80$ en el peor caso. Cabe mencionar en este análisis que excepto para $C50$ y $C80$, para ninguno de los otros 3 parámetros se alcanza el 6 % de error. Examinando el error cometido en el conjunto de parámetros, se aprecia una tendencia a cometer más error en los parámetros que originalmente dependían del cálculo previo de la respuesta impulsiva de la sala, como son $RT60$, $C50$ y $C80$, caso aparte el de STI que muestra una precisión excelente en predicción. Esta dificultad para estimar estos parámetros a partir de una señal de habla, que parece fallar para STI está motivada principalmente por la distribución y varianza de los datos de entrenamiento. Recordando la Figura 3.12 vemos que para $C50$ y $C80$ disponemos de datos peor distribuidos que para $RT60$, con muchos valores de estos parámetros con poca o nula representación en la base de datos. La distribución y varianza de los datos mejora considerablemente para STI, viéndose reflejado en los excelentes resultados de predicción, como sucede con SII, que si bién es un parámetros que sí se calcula originalmente a partir de una señal de habla, se puede observar que posee la distribución más constante de datos pese a tener poca representación de valores bajos de SII (inferiores a 0.3).

Los resultados de MRE que nos da una visión general del porcentaje de error pueden dar lugar a confusión, pues estamos trabajando con parámetros acústicos de magnitudes muy diferentes, donde para valores próximos a 0, por ejemplo, que son poco habituales en $C50$, un error de 1 dB representaría un porcentaje muy elevado, pero en la práctica sería prácticamente inapreciable. Por ello lo hemos complementado con el MAE, que nos muestra para cada parámetro el error medio cometido en las mismas unidades del parámetro estimado, por lo que es sumamente útil para relacionarlo con medidas prácticas reales.

Viendo los resultados del MAE podemos apreciar que de media por ejemplo en el caso de $RT60$ se comete un error de 0.065 s o lo que es lo mismo 65 ms que es una cifra totalmente aceptable para la mayoría de aplicaciones de monitorización de espacios acústicos siendo muy próximo a la mínima diferencia perceptible, como se comenta más adelante. Algo similar pasa con el error cometido en $C50$ y $C80$, donde el error no llega a 0.3 dB de media, lo que sería prácticamente inapreciable para el oído humano y ratifica el buen desempeño de la CNN a la hora de predecir los parámetros. Mención a parte tiene la predicción de los parámetros relacionados directamente con la inteligibilidad, STI y SII, donde el error cometido es el menor del 2.3 % en ambos casos. Si observamos el MAE cometido en la predicción de estos 2 parámetros, ronda el 0.02 de media, que para dos parámetros cuyos valores están dentro del intervalo [0 - 1] representa una precisión muy elevada.

Estos buenos resultados respecto a la precisión se ven ratificados observando los JND de los parámetros acústicos predichos por la CNN. Los JND, de *Just-Noticeable Difference* evalúan la sensación subjetiva que produce cada parámetro acústico y cuantifican la mínima diferencia que es capaz de percibir el ser humano promedio. Observando diferentes estudios [102, 103, 75], y para los parámetros de los que hay datos, los JND establecidos para los parámetros $RT60$, $C50$, $C80$ y STI se pueden ver en la Tabla 3.14. Junto a los JND incluimos el MAE y la desviación estándar del error cometido en predicción sobre el conjunto de datos de test por nuestro modelo de CNN. Observamos fácilmente que prácticamente en todos los casos estamos por debajo o muy por debajo del JND definido para el parámetro, pues únicamente para $RT60$ superamos en un 0.5 % el JND definido y una vez más, en un caso práctico, arrojaría resultados perfectamente útiles.

Tabla 3.14: JND y error en predicción, datos de Test.

	RT60	**C50**	**C80**	**STI**
JND	5 %	1 dB	1 dB	0.03
MAE	0.06	**0.28 dB**	**0.29 dB**	**0.01**
MRE (%)	**5.59**	7.82	9.29	2.19
Error σ	0.09	0.36	0.37	0.01

Queda patente por lo tanto el buen comportamiento en predicción, pues el error cometido, que por supuesto está presente, en la mayoría de aplicaciones sería asumible o inapreciable si observamos los JND para los distintos parámetros. Otro indicador del buen desempeño del modelo de CNN entrenado es el coeficiente de correlación cruzada ρ por parámetro que se puede ver en la anterior Tabla 3.13, con una media de 0.99 para los 5 parámetros predichos, y que sólo baja de 0.99 para SII con una correlación de 0.96, que es muy elevada también.

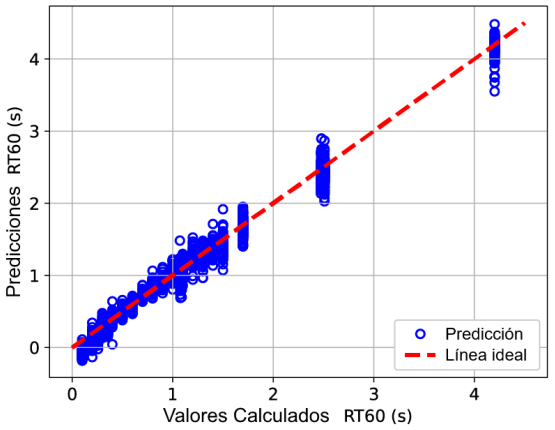
Distribución del error.

Además de los valores medios de error para cada parámetro acústico, es interesante observar cómo se distribuye el mismo en las predicciones. Para ello, en la Figura 3.16 se muestran los diagramas de dispersión con la predicción de los parámetros psicoacústicos $RT60$, $C50$, $C80$, STI y SII como círculos azules frente a los valores calculados de los mismos como líneas discontinuas rojas, para las señales del conjunto de datos de test.

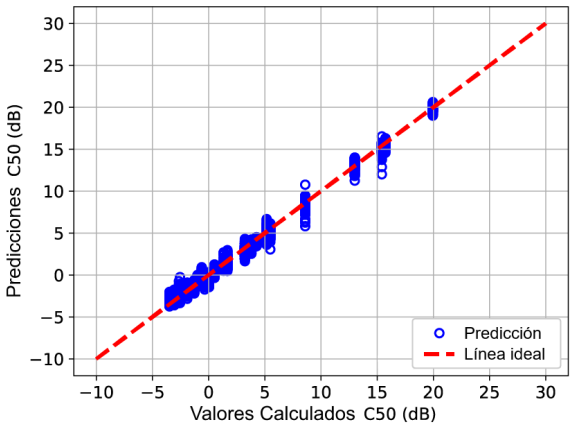
Se puede ver cómo la distribución de las predicciones se intenta ajustar a la línea ideal, alejándose lógicamente en zonas donde se disponía de menos datos en el entrenamiento, como por ejemplo para valores superiores a 2.5 s para $RT60$ (Figura 3.16a) o inferiores a 0.3 para SII (Figura 3.16e).

Se aprecia también como para los parámetros $C50$ y $C80$, donde el error medio es mayor que en el resto, las predicciones no se alejan tanto como podría parecer a priori, tal y como comentamos antes, debido a que el intervalo de valores para estros parámetros va de -10 dB a 40 dB y los errores situados próximos a 0 dB incrementan mucho la media porcentual de error, falseando el buen comportamiento de la CNN.

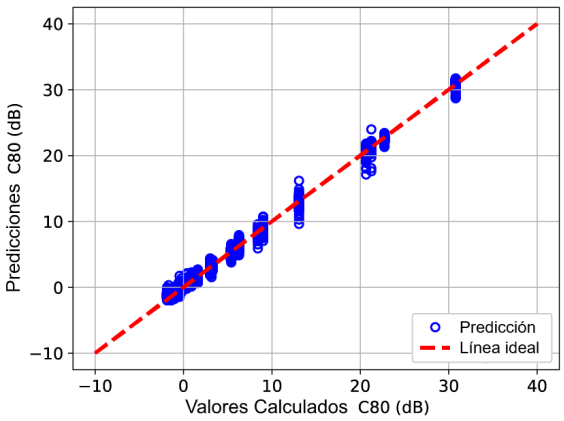
Si bien es habitual apreciar un efecto o error de sesgo en los diagramas de dispersión como los mostrados, en forma de valores de predicción que se alejan más de la recta ideal en los extremos, como se puede apreciar ligeramente en las predicciones de STI entre los valores de 0.3 y 0.4 o de SII de 0.2 a 0.4, este efecto sólo se nos presentó con las primeras configuraciones más sencillas de la CNN. Durante el proceso de diseño comprobamos que este efecto se mitigaba a medida que incrementábamos el número de parámetros entrenables de la red, aumentando el número de capas convolucionales y ajustando la dimensión de las mismas. Debido a esto no ha sido necesario tomar medidas para atenuar este efecto. Pese a las limitaciones expuestas, al observar la distribución de las predicciones de cada parámetro deducimos que en una aplicación real de la CNN, funcionando embebida en el sistema IoT de monitorización presentado en el apartado 2.2.2, los valores predichos para los parámetros acústicos de sala serían perfectamente válidos para la mayoría de análisis de sala rápidos donde más que la precisión se busca una idea general del comportamiento acústico de un recinto, como se mostrará más adelante.



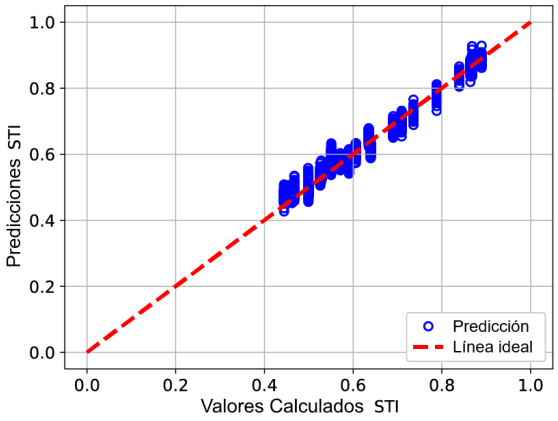
(a) $RT60$



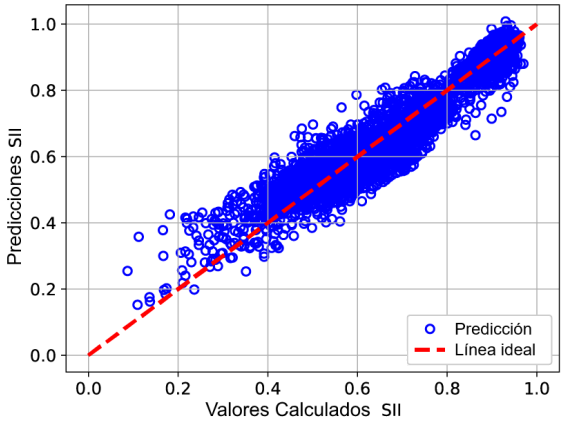
(b) $C50$



(c) $C80$



(d) STI



(e) SII

Figura 3.16: Parámetros acústicos de sala. Predicción vs. valores calculados.

Comparativa con estudios similares.

No existen demasiados estudios centrados en modelos de CNN capaces de predecir simultáneamente más de un parámetro acústico de sala. Menos aun que empleen técnicas de aprendizaje profundo para alimentar la red neuronal con audio sin procesar y no con parámetros del audio precalculados. Algunas de las aproximaciones existentes enfocadas a predecir parámetros acústicos de sala, lo hacen de forma individual (las más habituales) y no en conjunto como nuestro modelo. En este contexto, los modelos descritos en los estudios [90, 98]

y [99] han sido diseñados para predecir $RT60$, $C50$ y STI respectivamente. Respecto a modelos de CNN diseñados para predecir más de un parámetros acústico de forma simultánea, únicamente hemos hallado los estudios [100] y [101]. En concreto, [100] se centra en estimar 2 parámetros, $RT60$ y DRR, mientras que el estudio presentado en [101] considera un conjunto más amplio que incluye $RT60$, $C80$ y STI.

Debido a que estos estudios emplean bases de datos diferentes entre sí y diferentes a la nuestra también, la comparación de resultados directa no es precisa y se debe observar únicamente a título orientativo. Así, en una primera comparativa hemos empleado los resultados publicados en cada caso, recalculando nuestras métricas de error en términos de RMSE para contrastar resultados. En la Tabla 3.15 se muestran los valores publicados de RMSE y coeficiente de correlación cruzada para los trabajos mencionados junto con los obtenidos por nuestro modelo sobre el conjunto de datos de test.

Tabla 3.15: Comparativa de rendimiento con otros modelos de CNN.

	RT60		**C50**		**C80**		**STI**	
Ref. Modelo	**RMSE**	ρ	**RMSE**	ρ	**RMSE**	ρ	**RMSE**	ρ
[90] *	0.196	0.836	-	-	-	-	-	-
[98] *	-	-	3.300	0.840	-	-	-	-
[99] *	-	-	-	-	-	-	0.037	No data
[100] **	0.161	0.919	-	-	-	-	-	-
[101] **	0.393	0.918	-	-	2.038	0.943	0.040	0.913
Nuestro **	**0.115**	**0.995**	**0.458**	**0.996**	**0.472**	**0.997**	**0.017**	**0.991**
*	Modelos que predicen un solo parámetro							
**	Modelos que predicen más de un parámetro							

Tal y como mencionábamos, aunque la tabla puede ser útil para confirmar que los valores alcanzados se ajustan a los del estado del arte, con mejor rendimiento en todos los casos, hay que tener especial cuidado al comparar dichos valores debido a las consideraciones anteriores, pues la evaluación se está realizando sobre datos distintos.

Para poder realizar comparaciones precisas necesitamos o emplear los mismos datos que usa el estudio con el que queramos comparar sobre nuestro modelo de CNN, o implementar el modelo de CNN descrito en el estudio y evaluarlo con nuestros datos. Implementar los modelos descritos plantes el problema de cometer errores a la hora de definir las opciones de entrenamiento junto a otros hiper-parámetros de los modelos, debido a que no se describen con detalle en las publicaciones, por lo que optamos por buscar las bases de datos disponibles para emplearlas con nuestro modelo de CNN y evaluar el rendimiento.

Así pues, analizando las bases de datos empleadas en los mencionados estudios, procedimos a descartar las que no se adapten a nuestro caso, como las creadas a medida por no ser de acceso público y las que no se basan en señales de voz como entrada de los modelos. Eso nos deja con las empleadas en los estudios [90, 100, 101] y aunque nos interesó el caso de [101] debido a que el modelo descrito permite predecir más de 2 parámetros, no pudimos obtener la base de datos *SMILE-2004-Sound library of architecture and environment*. Debido a esto hemos seleccionado la base de datos del ACE Challenge [104], que se emplea tal cual en [90] y parcialmente en [100], además de haber sido sugerida por algunos revisores de la revista *IEEE-Internet Of Things Jounal* para realizar una evaluación más precisa de nuestro modelo.

La base de datos del ACE Challenge Corpus la conforman 3 elementos: diferentes respuestas impulsivas medidas en 7 salas diferentes, señales de ruido de 3 fuentes diferentes (ambiente, ventilador y murmullo) y señales de voz con frases tomadas de la base de datos TIMIT [85]. Empleando estos componentes, el software incluido en el ACE Challenge Corpus permite obtener 2 datasets, *Development* y *Evaluation*, con 2000 y 16000 señales respectivamente. Debido a que en el artículo de la referencia [100] emplean las respuestas impulsivas del ACE Challenge Corpus para probar técnicas de aumento de datos en el modelo de CNN, en lugar de los 2 conjuntos de datos tal cual, hemos decidido realizar la comparación con el estudio de la referencia [90], que es de los mismos autores pero que emplea los datos del ACE Challenge sin alterar. Empleando por lo tanto el conjunto de datos del ACE Challenge, hemos realizado diferentes pruebas para proporcionar una comparación más robusta frente a otro modelo de CNN empleando los mismos datos.

En la Tabla 3.16 se pueden ver los resultados de MSE y de correlación obtenidos al evaluar nuestro modelo frente a los publicados en [90] con los mismos datos del ACE Challenge Dataset. En primer lugar, hemos empleado el conjunto de datos *Evaluation* del ACE Challenge para probar nuestro modelo CNN sin reentrenamiento, obteniendo un valor de MSE de 0,041 en la predicción de RT60, que lógicamente es superior al obtenido sobre nuestro conjunto de datos de test.

Tabla 3.16: Evaluación del modelo con la base de datos del ACE Challenge.

	RT60	
Modelo de CNN	**MSE**	ρ
Ref. [90]	0.0384	0.836
Nuestro, sin reentrenamiento	0.0421	0.8101
Nuestro, con reentrenamiento	0.0309	0.9316

No obstante, pese a que este valor sitúa a nuestro modelo de CNN cerca del nivel de otros enfoques del estado del arte, nuestro modelo no ha contado en el proceso de entrenamiento con ningún de los incluidos en el ACE Challenge, lo que le permitiría adaptar aun más su respuesta. Para verificarlo, en segundo lugar, volvimos a entrenar nuestra red añadiendo el conjunto de datos *Development* del ACE Challenge Corpus a nuestros datos, sin cambiar ninguno de los hiperparámetros de la red neuronal, ni de diseño, ni de optimización, ni por supuesto de entrenamiento, para no cambiar el modelo diseñado y descrito en esta Tesis y poder evaluar su capacidad de adaptación únicamente añadiendo datos de naturaleza distinta a los empleados inicialmente. El modelo reentrenado proporciona mejores resultados de MSE al probarlo de nuevo con el conjunto de datos *Evaluation* del ACE Challenge, alcanzando un MSE de 0,030 y un mayor coeficiente de correlación también.

Es de remarcar llegados a este punto, que el rendimiento alcanzado por nuestro modelo de CNN es ligeramente superior al obtenido en [90] en términos de *RT60*, teniendo en cuenta que nuestro modelo propuesto está diseñado para proporcionar al mismo tiempo la estimación de 4 parámetros acústicos más simultáneamente, confirmando así objetivamente la eficacia del mismo. Además se ha podido comprobar como la inclusión de señales con diferentes características espectro-temporales permite a la CNN adaptarse al nuevo conjunto, sin haber cambiado parámetros de entrenamiento, por lo que deducimos que ajustando estos parámetros conseguiríamos una mejor respuesta aun.

Tiempo de procesado.

En el caso de los parámetros acústicos de la sala, puede parecer que no es demasiado importante poder trabajar en tiempo real, dado que las características físicas de la sala no cambian de manera brusca en el tiempo. Sin embargo, cambios en la ocupación de la sala o en el mobiliario de la misma pueden ocasionar que las características acústicas de esta varíen produciendo efectos indeseados. Por ejemplo, si deseamos medir la inteligibilidad del habla o la claridad en diferentes posiciones de una sala de forma continua mientras se pronuncia un discurso o una conferencia, es muy importante poder obtener estos parámetros lo más rápido posible para asegurar una buena comunicación.

La velocidad de cálculo es esencial también en el caso de una red WASN que implemente el cálculo en el propio nodo, ya que estos dispositivos, como hemos descrito anteriormente, disponen de capacidades de cálculo limitadas. Este es precisamente nuestro caso, ya que deseamos realizar el cálculo de los parámetros acústicos dentro de cada nodo del sistema IoT de monitorización que hemos diseñado, descrito en el apartado 2.2.2.

Así pues, para realizar esta prueba, hemos tomado el conjunto de datos de test con 1000 señales de voz y hemos medido el tiempo necesario para calcular los parámetros acústicos de sala directamente, a partir de las respuestas al impulso en el caso de $RT60$, $C50$, $C80$ y STI y a partir de la propia señal de voz en el caso del SII. A continuación, hemos medido el tiempo empleado por la CNN en predecir los mismos parámetros, en este caso sólo a partir de las señales de habla del mismo conjunto de datos. Para el proceso de cálculo directo mediante algoritmos de procesado de señal clásico, hemos empleado los algoritmos más optimizados, empleados ya para evaluar los tiempos de cálculo directo mostrados anteriormente en la Tabla 2.8 del apartado 2.4.2. Hemos utilizado también los mismos dispositivos que en el caso del cálculo directo, 2 ordenadores personales (1 de sobremesa y uno portátil) y 2 SBC, etiquetados como 'PC-1', 'PC-Portátil', 'UDOO X86' y 'Raspberry Pi', con las especificaciones técnicas mostradas anteriormente en la Tabla 2.2 de la sección 2.4, y que se resumen a continuación:

- PC-1: Intel(R) Core(TM) i7-7700 CPU @ 3.60 GHz, 32 GB RAM, 1 TB HDD.
- PC Portátil: Intel(R) Core(TM) i7-1065G7 CPU @ 1.3 GHz, 16 GB RAM, 500 GB SSD.
- Udoo X86 II-Ultra: Intel(R) Pentium (TM) N3710 @ 2.56 GHZ, 8 GB RAM, 500 GB HDD.
- Raspberry PI 3B: Broadcom BCM2837 CPU @ 1.2 GHz, 1 GB RAM, 16 GB SDHC class 10.

En la Figura 3.17 se puede ver de forma gráfica una comparativa del tiempo empleado por el cálculo directo frente al tiempo empleado por la CNN, para obtener los 5 parámetros acústicos de sala de nuestro conjunto, ($RT60$, $C50$, $C80$, STI y SII). Los valores mostrados son las medias obtenidas empleando los 1000 audios del conjunto de datos de test y están mostradas empleando escala logarítmica en el eje X, debido a la gran diferencia entre los tiempos obtenidos.

Probando sobre el ordenador de sobremesa (PC-1), el modelo de CNN permite predecir los parámetros acústicos de sala en un tiempo medio de 0.0039 s, frente a los 0.9590 s que emplea el cálculo directo, demostrando que la predicción mediante nuestro modelo es 245.8 veces más rápida. En el ordenador portátil, la relación disminuye ligeramente hasta 211.3 veces, manteniendo la ventaja no obstante de la red neuronal. Como se ha indicado, además de la evaluación sobre ordenadores personales, la buena precisión de la CNN la convierte en una opción viable para para ser empleada en dispositivos SBC en aplicaciones AI-IoT, por lo que es de especial interés su evaluación empleando estos.

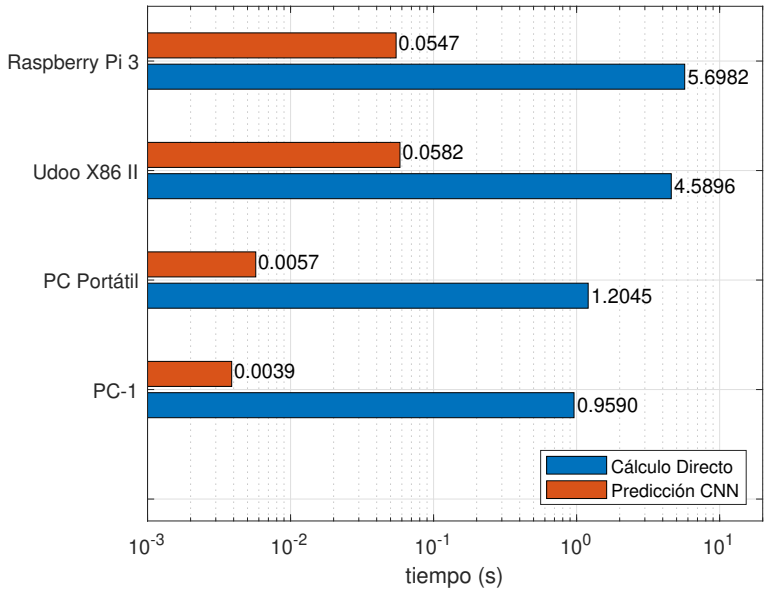


Figura 3.17: Parámetros de sala. Tiempos de cálculo directo vs predicción por CNN en diferentes plataformas.

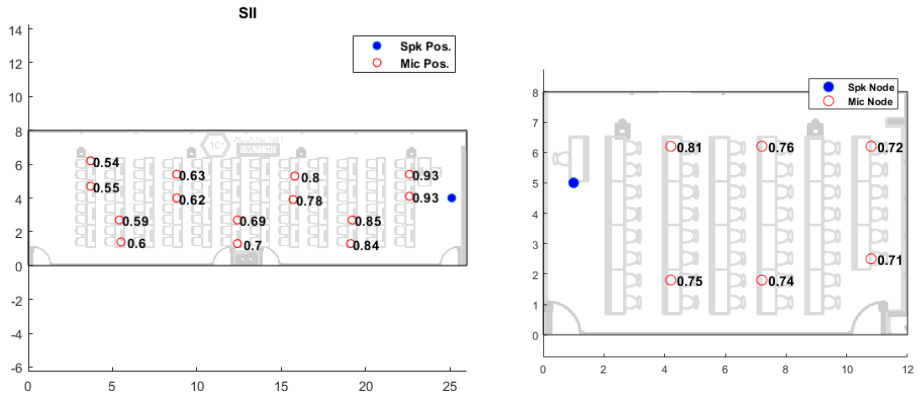
Debido a sus menores prestaciones, las pruebas sobre estos dispositivos presentan tiempos más elevados, pero manteniendo una clara ventaja para la predicción basada en nuestra CNN. Sobre la placa Udoo X86 la predicción es 78.8 veces más rápida que el cálculo directo, y utilizando la Raspberry Pi 3B llega a ser 104.1 veces más rápida. Este aumento en la velocidad de obtención de los parámetros acústicos de sala muestra la superioridad en velocidad de cálculo del modelo de CNN frente al cálculo directo, más teniendo en cuenta que la obtención de la respuesta impulsiva de la sala, necesaria para calcular ciertos parámetros, añade complejidad y tiempo de cálculo.

Así mismo, el sistema IoT con el modelo de CNN embebido permitiría la monitorización continua de un recinto, como un aula docente por ejemplo, mientras se desarrolla la actividad habitual en él sin tener que interrumpirla, gracias a que la predicción se realiza a a partir de señales de voz. Por otra parte, este aumento de velocidad de cálculo supone un importante ahorro de energía, permitiendo al sistema IoT emplear baterías y funcionar durante más horas en ausencia de conexión a la red eléctrica, lo que simplifica aun más la ubicación de los nodos del sistema.

Evaluación de RMSE en pruebas de campo.

Finalmente, dado el buen desempeño respecto a precisión y tiempo de procesado, hemos incluido el modelo de CNN entrenado en los nodos del sistema IoT descrito anteriormente en el apartado 2.2.2, conformando en definitiva un sistema AI-IoT que hemos empleado para realizar diferentes pruebas de campo en salas reales.

Estas pruebas tienen como objetivo realizar una prueba de concepto del marco total propuesto en esta parte de la Tesis Doctoral, evaluando el rendimiento del sistema AI-IoT completo en algunas aulas reales de las instalaciones de la Escuela Técnica Superior de Ingeniería de la Universidad de Valencia (ETSE-UV). Más concretamente, con estas pruebas se pretende validar la capacidad del modelo de IA para predecir con exactitud el valor real de los 5 parámetros acústicos considerados, realizando las predicciones dentro de los nodos del sistema IoT, transmitiendo y almacenando los valores calculados para su representación gráfica.



(a) Aula 1.0.1 (b) Aula 2.1.1

Figura 3.18: Ejemplo de posicionamiento de nodos del sistema AI-IoT con valores de *SII*.

El sistema AI-IoT probado se compone de 6 nodos receptores y 1 nodo de control, colocados como se muestra en la Figura 3.18 para el caso de las aulas 1.0.1 y 2.1.1, marcados con círculos rojos y azules respectivamente. En salas de grandes dimensiones o donde se deseen más puntos de muestreo, sólo es necesario reubicar los nodos y repetir el análisis, como se puede ver en la Figura 3.18a. Las fuentes de sonido han sido ubicadas en posiciones similares a las empleadas por los oradores en estos espacios. En todos los casos, para calcular los valores de referencia reales de los diferentes parámetros, se realizaron previamente mediciones de la respuesta impulsiva de la sala en las mismas posiciones en las que se colocaron los nodos del sistema IoT.

En la Figura 3.19 se puede ver el caso concreto del análisis realizado en el aula 1.1.3, con la posición de los nodos sobre el mapa de calor generado con los valores de inteligibilidad del habla *SII* predichos en cada posición de análisis. Hemos incluido en este caso la dirección a la que apunta el altavoz empleado por la directividad que puede presentar, como ayuda a la hora de interpretar la información.

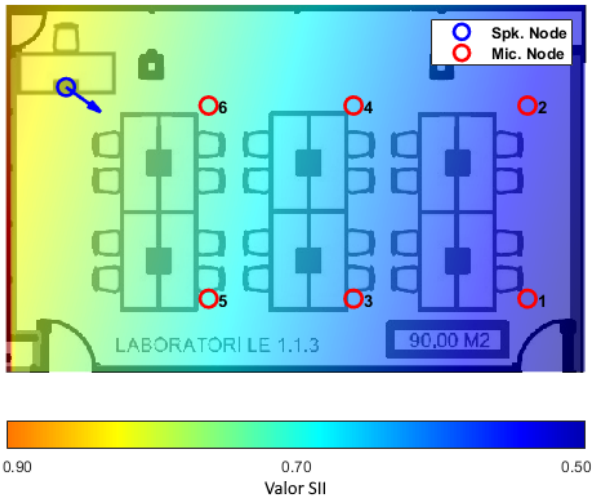


Figura 3.19: Prueba de campo en en el aula 1.1.3 de la ETSE-UV, mapa de calor de *SII* y ubicación de los nodos.

La Tabla 3.17 muestra los resultados obtenidos del análisis realizado en el aula 1.1.3, con los valores calculados frente a los estimados por la CNN para cada posición y para cada parámetro acústico. Para realizar las predicciones se han reproducido señales de habla distintas de las empleadas en las otras pruebas, extraídas de la base de datos *DARPA-TIMIT*

Tabla 3.17: Resultados de prueba de campo del sistema AI-IoT sobre el aula 1.1.3.

Nodo	**RT60 (s)**		**C50 (dB)**		**C80 (dB)**		**STI**		**SII**	
IoT	Calc.	$\overline{Pred.}$ (σ)	Calc.	$\overline{Pred.}$ (σ)	Calc.	$\overline{Pred.}$ (σ)	Calc.	$\overline{Pred.}$ (σ)	Calc.	$\overline{Pred.}$ (σ)
1	0.85	0.84 (0.10)	3.56	2.81 (1.10)	6.53	5.63 (1.28)	0.64	0.64 (0.03)	0.74	0.66 (0.06)
2	0.84	0.80 (0.08)	3.04	2.56 (1.05)	6.18	5.39 (1.24)	0.65	0.64 (0.02)	0.73	0.64 (0.06)
3	0.77	0.91 (0.12)	3.52	3.28 (1.18)	7.06	6.25 (1.40)	0.65	0.65 (0.02)	0.73	0.71 (0.08)
4	0.81	0.86 (0.12)	3.27	3.64 (1.02)	6.85	6.60 (1.19)	0.66	0.66 (0.02)	0.75	0.72 (0.06)
5	0.81	0.82 (0.10)	3.28	3.27 (1.03)	6.15	6.15 (1.24)	0.66	0.65 (0.02)	0.76	0.71 (0.06)
6	0.79	0.83 (0.11)	4.30	4.24 (1.05)	7.44	7.28 (1.25)	0.67	0.67 (0.02)	0.77	0.77 (0.06)
MAE		0.0483		0.3183		0.4850		0.0033		0.0450

acoustic-phonetic continuous speech database, grabando en cada nodo a 44100 Hz, remuestreando a 16 KHz y recortando la señal al tamaño de la entrada de la CNN. El proceso se ha repetido 10 veces para promediar varias predicciones y obtener una medición más fiable, por lo que en la Tabla 3.17 se muestra la media de estas 10 predicciones en la columna $\overline{Pred.}$ (σ), junto a la desviación estándar del error cometido entre paréntesis. La última fila de la tabla muestra el error medio absoluto de predicción de los 6 nodos para cada parámetro acústico.

Como era de esperar, el error de predicción es algo mayor que el obtenido al evaluar el conjunto de datos de test. Esto se debe probablemente al desajuste al que se enfrenta el modelo, que ahora predice sobre señales captadas dentro de una sala que no sólo no ha participado en el proceso de entrenamiento sino que además pueden presentar características no deseadas como el ruido propio del micrófono, el ruido de fondo ambiental u otros sonidos superpuestos que causan interferencias, debido a que el micrófono empleado es de bajo coste y a que no se realiza ningún procesado de la señal acústica para mitigar estos efectos. No obstante, viendo el rendimiento global del sistema, resumido en el MAE para los 6 nodos en la última fila de la Tabla 3.17, este se sitúa todavía por debajo de los valores JND [75, 102]. Los errores de predicción para $C50$, $C80$ y SII son un poco más altos y pueden deberse a los efectos no deseados comentados anteriormente, sin embargo los errores para $RT60$ y STI son ligeramente inferiores en esta prueba de campo a los de la evaluación con los datos de test.

Como se comentaba anteriormente, en la Figura 3.19 se muestra el mapa de calor generado con los valores de inteligibilidad SII estimados por el sistema, que en este caso es un parámetro muy dependiente de la distancia a la fuente de sonido. Pese a que el error es ligeramente superior al evaluado con los datos de test, este es de 0.045 de media, y sobre la escala de SII, que recordemos va de 0 a 1, representa un 4 % de la misma. Esto lo convierte en perfectamente útil para realizar cualquier análisis rápido, permitiendo detectar posibles problemas de inteligibilidad, como se ve gráficamente en la Figura 3.19 donde esta disminuye visiblemente al aumentar la distancia al orador, aunque en el caso del aula 1.1.3, todos los valores son superiores a 0.6, siendo buena inteligibilidad en líneas generales.

En cualquier caso, las predicciones obtenidas en esta prueba de campo, que es representativa de todas las realizadas en las diferentes aulas, han sido razonablemente precisas para nuestra aplicación, computacionalmente eficientes y libres de procedimientos de medición que implican la reproducción de señales de prueba específicas. Esto confirma la validez del marco propuesto para empleando nuestro modelo de CNN para la monitorización AI-IoT de los parámetros acústicos y de inteligibilidad del habla en la sala, permitiendo obtener una descripción significativa del comportamiento acústico de una sala.

Los resultados descritos en este apartado se han publicado con detalle en el artículo de revista [27], incluido en el compendio de publicaciones de esta Tesis como Anexo D.

Capítulo 4

Publicaciones y Contribuciones

En este capítulo se presentan las publicaciones realizadas durante el desarrollo de esta Tesis Doctoral y las contribuciones resultado de las mismas.

4.1. Publicaciones

La presente Tesis Doctoral ha sido financiada mediante la Ayuda para contratos predoctorales BES-2017-082340, financiada por el Ministerio de Ciencia e Innovación, la Agencia Estatal de Investigación, (MCIN/AEI /10.13039/501100011033) y por "FSE invierte en tu futuro". Esta ayuda está asociada al proyecto de investigación BIA2016-76957-C3-1-R, "Urbauramon, Herramientas inteligentes para la gestión y control del paisaje sonoro urbano. Definición de protocolos de monitorización y auralización. Intervención en el Patrimonio Sonoro", correspondiente al Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad, en el Marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016. Ello ha propiciado la colaboración con personal investigador multidisciplinar de diferentes entidades.

El trabajo de investigación desarrollado en la presente Tesis Doctoral ha dado lugar a diferentes resultados que se han publicado en revistas indexadas en el índice internacional *Journal Citation Reports* (JCR) y presentado a diversos congresos nacionales e internacionales. Estas publicaciones se han realizado en coautoría con los directores de la presente Tesis además de distinto personal docente e investigador del departamento de Informática y del departamento de Fisiología de la Universidad de Valencia, del departamento de Física Aplicada de la Universidad Politécnica de Valencia, del departamento de Física de la Universidad Pública de Navarra, de la Escuela Técnica Superior de Arquitectura de Barcelona de la Universidad Politécnica de Cataluña Barcelona Tech y de la *School of Computing, Engineering and Physical Sciences* de la *University of the West of Scotland*, con los que se ha colaborado a lo largo del desarrollo de esta investigación.

4.1.1. Publicaciones en revistas

Detalle de artículos publicados en revistas:

- Autores: Jesus Lopez-Ballester, Adolfo Pastor-Aparicio, Jaume Segura-Garcia, Santiago Felici-Castell and Maximo Cobos.

Título: Computation of psycho-acoustic annoyance using deep neural networks
Año: 2019
Revista: Applied Sciences-Basel, Vol. 9, Nº15, Pags. 1-12
Factor de impacto: 2.474 (Q2)
DOI: 10.3390/app9153136. [24]

- Autores: A. Pastor-Aparicio, J. Segura-Garcia, J. Lopez-Ballester, S. Felici-Castell, M. Garcia-Pineda, J.J. Pérez-Solano
  Título: Psycho-Acoustic Annoyance Implementation with Wireless Acoustic Sensor Networks for Monitoring in Smart Cities
  Año: 2020
  Revista: IEEE - Internet Of Things Journal, Vol. 7, Nº 1, Pags. 128-136
  Factor de impacto: 9.471 (Q1)
  DOI: 10.1109/JIOT.2019.2946971. [83]

- Autores: Jesus Lopez-Ballester, Adolfo Pastor-Aparicio, Santiago Felici-Castell, Jaume Segura-Garcia, Maximo Cobos
  Título: Enabling real-time computation of psycho-acoustic parameters in acoustic sensors using convolutional neural networks
  Año: 2020
  Revista: IEEE - Sensors Journal, Vol. 20, Nº 19, Pags. 11429-11438
  Factor de impacto: 3.301 (Q2)
  DOI: 10.1109/JSEN.2020.2995779. [25]

- Autores: Jesus Lopez-Ballester, Jose M. Alcaraz Calero, Jaume Segura-Garcia, Santiago Felici-Castell, Miguel Garcia-Pineda, Maximo Cobos
  Título: Speech Intelligibility Analysis and Approximation to Room Parameters through the Internet of Things
  Año: 2021
  Revista: Applied Sciences-Basel, Vol. 11, Nº 4, Pags. 1430-1440
  Factor de Impacto: 2.838 (Q2)
  DOI: 10.3390/app11041430. [26]

- Autores: Jesus Lopez-Ballester, Santiago Felici-Castell, Jaume Segura-Garcia, Maximo Cobos
  Título: AI-IoT Platform for Blind Estimation of Room Acoustic Parameters Based on Deep Neural Networks
  Año: 2022
  Revista: IEEE - Internet Of Things Journal
  Factor de Impacto (año 2021): 10.238 (Q1)
  DOI: 10.1109/JIOT.2022.3203570. [27]

4.1.2. Publicaciones en congresos

Detalle de comunicaciones expuestas en congresos:

- Autores: Jaume Segura Garcia, Adolfo Pastor-Aparicio, Jesus Lopez-Ballester, Juan J. Perez-Solano, Santiago Felici-Castell, Maximo Cobos-Serrano, Francisco Grimaldo-Moreno, Miguel Arana-Burgui, Francesc Daumal-Domènech, Rosa Cibrián Ortiz de Anda, Alicia Giménez-Pérez
  Título: Descripción del paisaje sonoro de las Fallas de València (2). Percepción subjetiva y parámetros psicoacústicos
  Año: 2018

Congreso: FIA 2018 - XI Congreso Iberoamericano de Acústica; X Congreso Ibérico de Acústica; 49º Congreso Español de Acústica - TECNIACUSTICA'18, 24-26 de octubre. [105]

- Autores: Jaume Segura Garcia, Adolfo Pastor-Aparicio, Jesus Lopez-Ballester, Juan J. Perez-Solano, Santiago Felici-Castell, Máximo Cobos-Serrano, José Montoya-Belmonte, Ana Torres-Aranda, Juan M. Navarro-Ruiz
  Título: Análisis espacio-temporal de parámetros psico-acústicos en entornos acústicos urbanos usando sistemas IoT en tiempo real
  Año: 2018
  Congreso: FIA 2018 - XI Congreso Iberoamericano de Acústica; X Congreso Ibérico de Acústica; 49º Congreso Español de Acústica - TECNIACUSTICA'18, 24-26 de octubre. [106]

- Autores: Jaume Segura-Garcia; Jesus Lopez-Ballester; Adolfo Pastor-Aparicio; Santiago Felici-Castell; Máximo Cobos-Serrano; Juan J. Pérez-Solano; Antonio Soriano-Asensi; Miguel García-Pineda
  Título: Visualization of nuisance information in acoustic environments using an IoT system
  Año: 2019
  Congreso: 48th International Congress and Exhibition on Noise Control Engineering - INTER-NOISE 2019. [81]

- Autores: Adolfo Pastor-Aparicio, Jesus Lopez-Ballester, Jaume Segura-Garcia, Santiago Felici-Castell, Maximo Cobos-Serrano, Rafael Fayos-Jordn, Rosa Maria Cibrian, Alicia Gimenez-Perez, Miguel Arana-Burgui
  Título: Zwicker's Annoyance model implementation in a WASN node
  Año: 2019
  Congreso: 48th International Congress and Exhibition on Noise Control Engineering - INTER-NOISE 2019. [84]

- Autores: Adolfo Pastor-Aparicio; Jesus Lopez-Ballester; Santiago Felici-Castell; Jaume Segura-Garcia; Rafael Fayos-Jordán; Miguel Garcia-Pineda
  Título: Efficient implementation of an IoT deployment for sound-scape monitoring
  Año: 2019
  Congreso: XIV Jornadas de Ingeniería Telemática - JITEL 2019. [107]

- Autores: Adolfo Pastor-Aparicio; Jesus Lopez-Ballester; Santiago Felici-Castell; Jaume Segura-Garcia; Rafael Fayos-Jordán; Juan José Pérez-Solano; Máximo Cobos; Miguel Garcia-Pineda
  Título: URBAURAMON: Herramientas inteligentes para la gestión y control del paisaje sonoro urbano
  Año: 2019
  Congreso: XIV Jornadas de Ingeniería Telemática - JITEL 2019. [108]

- Autores: Jaume Segura-Garcia, Santiago Felici-Castell, Jose M. Alcaraz Calero, Qi Wang, Jesus Lopez-Ballester, Rafael Fayos-Jordán, Juan J. Pérez-Solano, Miguel Arana-Burgui

Título: Sistema basado en tecnología 5G de monitorización psicoacústica del paisaje sonoro en Smart Cities con offloading computacional dinámico en el Edge
Año: 2020
Congreso: Acustica 2020 - XI Congreso Ibérico de Acústica y 51º Congreso Español de Acústica -TECNIACUSTICA'20, 21-23 de octubre. [109]

- Autores: Jesus Lopez-Ballester, Rafael Fayos-Jordan, Jaume Segura-Garcia, Santiago Felici-Castell, Juan J. Pérez-Solano, Maximo Cobos, Rosa Cibrián, Alicia Giménez-Pérez
  Título: Análisis de inteligibilidad y aproximación a parámetros de sala mediante Internet de las Cosas
  Año: 2020
  Congreso: Acustica 2020 - XI Congreso Ibérico de Acústica y 51º Congreso Español de Acústica -TECNIACUSTICA'20, 21-23 de octubre. [79]

- Autores: Jesus Lopez-Ballester, Santiago Felici Castell, Jaume Segura-Garcia, Juan José Perez-Solano, Antonio Soriano-Asensi
  Título: URBAURAMON: Herramientas inteligentes para la gestión y monitorización acústica
  Año: 2021
  Congreso: XV Jornadas de Ingenieria Telematica - JITEL 2021. [110]

- Autores: Jesus Lopez-Ballester, Santiago Felici Castell, Jaume Segura-Garcia, Juan José Perez-Solano, Antonio Soriano-Asensi
  Título: Sistema IoT para la monitorización de parámetros de sala e inteligibilidad del habla
  Año: 2021
  Congreso: XV Jornadas de Ingenieria Telematica - JITEL 2021. [80]

- Autores: Jaume Segura-Garcia, Jesus Lopez-Ballester, Santiago Felici-Castell, Juan J. Perez-Solano, Jose M. Alcaraz-Calero, Rafael Fayos-Jordan, Enrique A. Navarro-Camba, Antonio Soriano-Asensi, Juan M. Navarro-Ruiz
  Título: Soundscape monitoring of modified psychoacoustic annoyance with Next-Generation EDGE computing and IoT
  Año: 2022
  Congreso: 11th Euro American Conference on Telematics and Information Systems - EATIS 2022. [111]

4.2. Contribuciones y Compendio de Publicaciones

Las publicaciones derivadas del trabajo desarrollado en esta Tesis pueden agruparse en 4 contribuciones que se describen a continuación.

4.2.1. Primera Contribución

Desarrollo y optimización de los algoritmos necesarios para el cálculo de parámetros de molestia psicoacústica incluidos en el modelo de Zwicker, N, S, R, FS. Desarrollo de un

sistema completo que permite obtener todos los parámetros mencionados además del indicador general de molestia psicoacústica PA. Diseño e implementación de un sistema IoT de monitorización de bajo coste que permite el cálculo de estos parámetros ya sea en los nodos distribuidos o en el nodo central, minimizando los efectos de las limitaciones que presentan los dispositivos de recursos limitados.

Implementación de un sistema que permite la visualización intuitiva de los valores de los parámetros calculados para cada posición estudiada mediante los nodos del sistema IoT.

Los elementos desarrollados en esta contribución se han expuesto en diferentes congresos nacionales e internacionales a lo largo de su evolución, como puede verse en las referencias [81, 84, 105, 106, 107, 108, 109, 111].

4.2.2. Segunda Contribución

Diseño y entrenamiento de una red neuronal convolucional que permite la predicción de los parámetros de molestia psicoacústica mediante aprendizaje profundo, con una precisión elevada y una carga computacional muy inferior al cálculo directo. Desarrollo de una base de datos propia basada en sonidos urbanos para realizar el entrenamiento y verificación del modelo de CNN diseñado.

Inclusión del modelo de CNN entrenado en el sistema de monitorización IoT permitiendo la monitorización y visualización de parámetros psicoacústicos en tiempo real.

Esta contribución se ha publicado en dos artículos de revista (Anexos A y B) que forman las dos primeras publicaciones del compendio que se presenta en esta Tesis:

- "Computation of psycho-acoustic annoyance using deep neural networks", Applied Sciences-Basel, 9 (15), 1-12, 2019. [24]
- "Enabling real-time computation of psycho-acoustic parameters in acoustic sensors using convolutional neural networks", IEEE - Sensors Journal, 20 (19), 11429-11438, 2020. [25]

4.2.3. Tercera Contribución

Desarrollo y optimización de los algoritmos necesarios para el cálculo de un conjunto de parámetros acústicos de sala, $RT60$, $C50$, $C80$, STI y SII. Implementación de un sistema completo que permite obtener la respuesta impulsiva de una sala y calcular el conjunto de parámetros acústicos empleando esta junto a señales de voz. Mejora del sistema IoT de monitorización diseñado mediante la inclusión del protocolo de comunicaciones MQTT y adaptación del mismo al cálculo de los parámetros acústicos de sala. Adaptación del sistema de representación implementado a los nuevos parámetros mejorando las características de la representación.

Esta contribución se ha expuesto en diferentes congresos nacionales e internacionales, como puede verse en las referencias [79, 80, 110]. Además, la contribución se ha publicado en un artículo de revista que forma la tercera publicación (Anexo C) del compendio que se presenta en esta Tesis:

- "Speech Intelligibility Analysis and Approximation to Room Parameters through the Internet of Things", Applied Sciences-Basel, 11 (4), 1430-1440, 2021. [26]

4.2.4. Cuarta Contribución

Diseño y entrenamiento de una red neuronal convolucional que permite la predicción de los parámetros de sala mencionados en el punto anterior mediante aprendizaje profundo, a

partir de señales de voz, evitando el proceso previo de cálculo de la respuesta impulsiva de la sala, con elevada precisión y un tiempo de procesado muy inferior al empleado por el cálculo directo empleado en el apartado anterior. Este proceso incluye el desarrollo de una base de datos propia basada en respuestas impulsivas tanto sintéticas como reales y señales de habla que permite el entrenamiento y verificación del modelo de CNN diseñado.

El rendimiento del modelo se ha evaluado empleando también una base de datos pública conocida, tanto sin reentrenar como reentrenando para ajustar la respuesta a los nuevos datos, devolviendo resultados muy buenos teniendo en cuenta que no se han modificado ni ningún hiperparámetro del modelo de CNN ni ninguna opción de entrenamiento.

Inclusión del modelo de CNN entrenado en el sistema de monitorización IoT permitiendo la monitorización y visualización de los parámetros acústicos de sala de manera rápida y precisa, verificado con pruebas de campo en entornos reales.

Esta contribución ha sido publicada en el artículo de revista que define la cuarta publicación del compendio de esta Tesis, incluido como Anexo D:

- "AI-IoT Platform for Blind Estimation of Room Acoustic Parameters Based on Deep Neural Networks", IEEE - Internet Of Things Journal, 2022. [27]

Capítulo 5

Conclusiones

En este capítulo que concluye la memoria de esta Tesis Doctoral, se exponen las conclusiones extraídas del desarrollo de la misma, además de las ideas relativas a trabajos futuros que se podrían realizar.

5.1. Conclusiones

Los parámetros acústicos nos permiten cuantificar de manera objetiva diferentes características tanto del sonido que nos rodea en un emplazamiento concreto, que define el paisaje sonoro o soundscape, en este caso mediante parámetros psicoacústicos, como del comportamiento acústico del entorno, mediante los parámetros acústicos de sala.

Los parámetros acústicos son dependientes y propios del emplazamiento donde se miden y del sonido presente en el mismo, por lo que disponer de un sistema de monitorización que permita estimarlos en diferentes posiciones es de sumo interés. Las redes de sensores acústicos inalámbricos o WASN pueden ser muy útiles en este cometido gracias a sus características crecientes en materia de almacenamiento, conectividad y bajo consumo, que ha propiciado la aparición del Internet de las Cosas o Internet of Things. Sin embargo existen ciertas dificultades para implementar el cálculo de los parámetros acústicos dentro de nodos IoT, por un lado relativas a la complejidad de los algoritmos de cálculo, y por otro a los diferentes procesos previos necesarios, como puede ser la obtención de la respuesta impulsiva de la sala. El desarrollo de nuevos sistemas de monitorización acústica, tiene que aportar soluciones a estas dificultades planteadas, tanto a nivel de aceleración de cálculos, como a nivel de simplificación de despliegues. La inteligencia artificial (AI) en forma de redes neuronales convolucionales, puede aportar soluciones a estos problemas dando lugar a un sistemas de monitorización AI-IoT que permitan obtener los parámetros de forma rápida y precisa, sin necesidad de complejas operaciones previas o grandes despliegues de sistemas cableados.

En esta Tesis Doctoral se presenta un sistema de monitorización AI-IoT que permite obtener tanto parámetros de molestia psicoacústica como parámetros acústicos de sala. Empleando algoritmos de cálculo clásico no se consigue sortear los problemas de velocidad y complejidadde cálculo planteadas por lo que se han desarrollado dos modelos de CNN que mediante aprendizaje profundo permiten predecir parámetros relativos a la molestia psicoacústica recogidos en el modelo de Zwicker y un conjunto de parámetros acústicos de sala, únicamente a partir de señales de audio sin procesar, sin necesidad de extraer respuestas impulsivas ni realizar operaciones previas de filtrado o extracción de características de las señales.

En la Sección 2.1 de esta memoria se describen los métodos para obtener tanto los parámetros acústicos recogidos en el modelo de molestia psicoacústica de Zwicker, como los parámetros acústicos de sala. Estos algoritmos se han optimizado buscando el menor tiempo de cálculo posible y se han incorporado al sistema de monitorización IoT diseñado, como se describe en la Sección 2.2.2. El diseño del sistema IoT y la incorporación de los algoritmos de cálculo directo al mismo forman dos contribuciones, una publicada como diversas ponencias de congreso (según su estado de desarrollo) y otra como el artículo de revista incluido en el Anexo C.

Del desarrollo del sistema IoT de monitorización que emplea algoritmos de cálculo directo podemos concluir que pese a que existen diversos sistemas de medición en el mercado, ninguno ofrece la versatilidad del propuesto ni la facilidad de despliegue a un coste tan reducido. Aunque los sistemas existentes presentan una complejidad técnica más elevada, no permiten la monitorización de diferentes puntos de forma simultánea como implementa el diseñado. La complejidad del sistema se ha reducido al mínimo para facilitar su uso y únicamente haya que emplazar los nodos para comenzar la monitorización, pues las comunicaciones se efectúan de forma inalámbrica mediante WiFi. Para ello, el nodo central del sistema IoT controla tanto la sincronización del sistema mediante un protocolo de comunicaciones como el almacén de los datos monitorizados, de forma transparente al usuario y permitiendo una sincronización con retrasos inferiores a los 100 uS, aunque en el caso actual no es necesaria una sincronización precisa, pues los cálculos se realizan con la señal de cada nodo individualmente. Al haber implementado un sistema de visualización a medida, la calidad de la información gráfica también es más elevada que en la mayoría de sistemas existentes, pudiendo ser configurada además a medida de cada necesidad. El sistema IoT de monitorización permite independizarse de la red eléctrica al funcionar con baterías integradas en los nodos, lo que facilita su despliegue y permite el funcionamiento durante horas, gracias también al reducido consumo.

A la hora de monitorizar parámetros de molestia psicoacústica, como conclusión extraemos que el sistema diseñado permitiría mediante un despliegue rápido y sencillo el análisis de diferentes soundscapes, como pueden ser un barrio de una ciudad o una zona verde y evaluar los niveles de contaminación acústica que afectan al confort y la habitabilidad, para actuar en caso de ser necesario. En el caso de los parámetros acústicos de sala, el sistema permite monitorizar en diferentes zonas de la misma para evaluar el comportamiento acústico, permitiendo detectar los posibles puntos conflictivos y actuar en consecuencia, permitiendo atenuar ciertos efectos indeseados en una sala de conciertos orientada a reproducción musical o amplificando otros en recintos destinados al transporte de pasajeros por ejemplo, permitiendo así adecuar la acústica de cada espacio a su uso final.

No obstante, como se ha descrito en la Sección 2.4, tanto para los parámetros psicoacústicos como para los parámetros acústicos de sala, el cálculo directo plantea una serie de problemas al ser implementado en los nodos del sistema IoT, pese a los esfuerzos realizados para solucionarlos empleando un enfoque de programación clásico. En esta Tesis Doctoral se ha planteado la solución a este problema mediante un enfoque basado en redes neuronales convolucionales entrenadas empleando aprendizaje profundo para realizar la predicción de diferentes parámetros acústicos según sea el caso, a partir de señales en crudo sin procesar.

En la Sección 3.2 se describe el diseño, entrenamiento y evaluación de un modelo de CNN destinado a la predicción de los parámetros psicoacústicos que forman el modelo de Zwicker: N, S, R, F y el indicador de molestia PA. Empleando como entrada señales de audio sin procesar y en un tiempo muy inferior al empleado por el cálculo directo. Esta contribución ha dado lugar a los 2 artículos de revista incluidos en los Anexos A y B. De ellos podemos concluir que la CNN propuesta ha demostrado ser muy precisa en el cálculo de los parámetros

psicoacústicos, especialmente en Loudness (N) y en el indicador de molestia general PA, con valores de error medio inferiores al 3 %. Para entrenar el modelo se ha empleado un amplio conjunto de datos de sonidos urbanos reales divididos en segmentos de 1 s de duración y pese a presentar cierto desequilibrio por no disponer del mismo número de datos para todo el rango de valores de los parámetros, el modelo resultante ha demostrado un buen desempeño. Se comete lógicamente menos error en los rangos donde se ha dispuesto de más datos para entrenar, como por ejemplo en valores de PA situados entre 0 y 50.

Por otra parte, comparando el tiempo requerido para obtener los parámetros psicoacústicos, el modelo propuesto logra una velocidad 250 veces superior al cálculo directo, así como un uso inferior de memoria RAM, elementos críticos para el funcionamiento en dispositivos SBC de bajo coste. Esto permite realizar los cálculos en tiempo real, gracias a que el tiempo requerido es ahora de sólo 0.02 s frente a los 1.52 s anteriores, para analizar 1 s de sonido, algo que es vital para no perder tramas de sonido mientras calculamos, pero lo que es más importante es que permite realizar la predicción dentro de cada nodo del sistema IoT, por lo que se libera al sistema de la transmisión de las señales de audio y al nodo central de la tarea de realizar los cálculos. Como consecuencia la escalabilidad del sistema, antes limitada por el número máximo de nodos que mandaban secuencias de sonido al nodo central, se incrementa en varios órdenes de magnitud, puesto que únicamente se transmiten al nodo central o en su defecto al servidor los valores calculados de los parámetros en cada nodo. Dado que no es necesario transmitir secuencias de audio, eludimos problemas de privacidad y con leyes de protección de datos, pues no se transmite ni almacena ninguna información de naturaleza sensible, habilitando el sistema para monitorizar emplazamientos cuya información acústica tenga naturaleza confidencial. Así pues, concluimos que se ha conseguido implementar un sistema AI-IoT de monitorización de parámetros psicoacústicos gracias a un enfoque basado en redes neuronales convolucionales y aprendizaje profundo, que posee una buena precisión y gran velocidad de cálculo, por lo que permite monitorizar soundscapes de dimensiones considerables en tiempo real y con gran facilidad de despliegue, permitiendo almacenar los datos de forma local o en la nube y representar los mismos a voluntad del usuario para su análisis. Llegados a este punto en las conclusiones, tras analizar el diseño del sistema AI-IoT que incorpora el modelo de CNN orientado al primer conjunto de parámetros, podemos verificar que se han cumplido los objetivos iniciales de esta Tesis Doctoral descritos en la Sección 1.2, al menos al 50 % que respecta a los parámetros de molestia psicoacústica.

Para verificar el cumplimiento del resto de objetivos, analizaremos la incorporación en el sistema AI-IoT de un modelo de CNN en este caso orientado a los parámetros acústicos de sala. En la Sección 3.3 se describe el diseño, entrenamiento y evaluación de un modelo de CNN orientado esta vez a la predicción de los parámetros acústicos de sala: $RT60$, $C50$, $C80$, STI y SII. El modelo de CNN está basado en aprendizaje profundo, por lo que emplea como señales crudas de voz que al han sido modificadas por las características constructivas de la sala a analizar, al haberse producido en esta. Para ello se ha creado una base de datos que combina tanto respuestas impulsivas reales, como sintéticas junto con señales de voz. De las numerosas pruebas descritas en esa sección, se puede concluir que el modelo de CNN diseñado presenta una precisión muy buena para la mayoría de aplicaciones, con un MRE cercano al 5 % de media e inferior al 3 % en parámetros relacionados con inteligibilidad como son STI y SII. Para verificar objetivamente la precisión del modelo de CNN más allá de nuestra aplicación, se ha comparado el error cometido con las diferencias mínimas perceptibles para cada parámetros (JND), obteniendo errores inferiores a los JND en todos los casos excepto para $RT60$, donde el error iguala al JND situándolo en el límite de la precisión perceptible. Por otra parte, para reforzar la evaluación objetiva de precisión del modelo de CNN diseñado, se ha evaluado empleando datos de otra base de datos reconocida (ACE Challenge database), situando su rendimiento al mismo nivel que el de otros enfoques

del estado del arte, si bien dedicados a predecir un solo parámetro y no 5 como nuestro modelo. Cabe mencionar que se ha evaluado la capacidad plástica del modelo para aprender de nuevos datos, logrando superar el rendimiento de otros enfoques únicamente incluyendo los nuevos datos en el proceso de entrenamiento, sin la modificación de ningún hiperparámetro ni de la configuración de entrenamiento.

De estas evaluaciones extraemos la conclusión de que la precisión alcanzada por el modelo propuesto es adecuada para su uso en un sistema de monitorización, como demuestran además las pruebas de campo que hemos realizado en salas reales, donde que se ha evaluado el modelo incluyéndolo en el sistema AI-IoT diseñado, gracias a la reducida carga computacional que implica.

Evaluando la velocidad de procesado del modelo de CNN respecto al cálculo directo a la hora de obtener los parámetros acústicos de sala, obtenemos un incremento de velocidad de más de 245 veces, y lo que es más importante, permitiendo realizar la predicción dentro de los mismos nodos del sistema IoT. Mediante el modelo de CNN diseñado y sobre un dispositivo típico SBC (Raspberry Pi) los 5 parámetros acústicos de sala se obtienen en una media de 50 ms, a partir de una señal de voz sin procesar de 3.5 segundos de duración, evitando el proceso de cálculo de la respuesta impulsiva con lo que implica una gran simplificación del proceso. Como en el caso de los parámetros psicoacústicos, se libera al sistema de la transmisión de ninguna señal acústica y al nodo central de realizar ningún cálculo, por lo que la escalabilidad del sistema vendrá definida por otros factores, como el número de dispositivos que soporte la red inalámbrica. En este punto, es de remarcar que se evitan problemas de privacidad y con leyes de protección de datos, pues únicamente se conserva y transmiten los valores de los parámetros acústicos de sala, que no recogen ninguna información de estos tipos. Se puede concluir por lo tanto que la predicción de los parámetros acústicos de sala mediante el enfoque basado en CNNs entrenadas mediante aprendizaje profundo es totalmente viable, como se ha demostrado con éxito, incorporando además el modelo diseñado al sistema de monitorización AI-IoT y permitiendo el cálculo dentro de los mismos nodos sin enviar ninguna secuencia de audio y lo que es de remarcar, sin necesidad de la extracción previa de la respuesta impulsiva de la sala.

Una conclusión clara de este aspecto es que poder obtener los parámetros acústicos de forma sencilla y rápida, mediante un sistema facil de desplegar y utilizar como el AI-IoT diseñado, permite detectar y actuar de manera inmediata frente a problemas acústicos en espacios donde la comunicación sea esencial. Si anteriormente se han mencionado espacios docentes y lugares de acceso público, cabe mencionar que en los últimos años está ganando protagonismo el análisis acústico de espacios médicos donde la inteligibilidad de la palabra es vital, como los quirófanos de hospital. Además, con la aparición de salas destinadas a intervenciones quirúrgicas de forma telemática, es necesario garantizar espacios con la respuesta acústica adecuada y este tipo de emplazamientos suelen presentar problemas para desplegar un sistema cableado complejo, por lo que el empleo de un sistema AI-IoT es una solución perfecta.

Esta contribución se ha publicado como un artículo de revista que a fecha de hoy se encuentra en fase de revisión, aceptado para publicación con cambios menores.

En definitiva, el proceso de investigación llevado a cabo en esta Tesis Doctoral ha partido del desarrollo y depuración de diferentes algoritmos destinados a obtener parámetros de molestia psicoacústica y parámetros acústicos de sala para poder implementar nuevos métodos de monitorización. Los problemas relativos al tiempo requerido por los cálculos del procesado de señal necesario han conducido a afrontar su obtención desde un enfoque basado en redes neuronales convolucionales basadas en aprendizaje profundo, obteniéndose modelos de CNN que han proporcionado la precisión y el rendimiento adecuados para su uso. Se ha demostrado como los modelos de CNN son capaces de extraer características de las señales acústicas,

tanto intrínsecas a ellas como relativas al entorno donde se reproducen, aprender de estas características extraídas y proporcionar mediante regresión predicciones precisas tanto de parámetros psicoacústicos como de parámetros acústicos de sala en un tiempo varios ordenes de magnitud inferior al empleado por el cálculo directo. Esto hace posible que sean ejecutados en dispositivos de bajo coste y recursos limitados. Paralelamente se ha desarrollado un sistema IoT de monitorización inalámbrica donde se han incorporado los modelos de CNN con éxito, y que permite la monitorización tanto de soundscapes exteriores amplios como de espacios interiores de manera rápida y sencilla.

Gracias a esto podemos concluir de forma global que el enfoque de la monitorización acústica mediante CNNs embebidas en un sistema AI-IoT, es una alternativa perfectamente viable a los métodos tradicionales de monitorización. El hecho de basar la monitorización en un sistema IoT permite el análisis de diferentes puntos en entornos amplios, que combinado con la velocidad de obtención de los parámetros y a diferentes representaciones numéricas y gráficas, permite una comprensión más profunda a la vez que sencilla de los análisis acústicos que se pueden llevar a cabo.

5.2. Trabajo Futuro

La investigación realizada en esta Tesis Doctoral ha abarcado muchos de los problemas que se plantean en la monitorización acústica de soundscapes y de recintos. Al mismo tiempo que se ha aportado las soluciones descritas en esta memoria con resultados muy satisfactorios, se han planteado diferentes cuestiones alternativas susceptibles de ser investigadas en nuevas líneas de investigación.

Las más significativas son las siguientes:

- **Incremento de la frecuencia de muestreo:** el empleo de una frecuencia de muestreo de 16 kHz vino impuesto desde un principio por la definición de muchos parámetros acústicos, junto con limitaciones de hardware tanto de monitorización como de cálculo y entrenamiento de las CNN. Esta limitación, mencionada en las contribuciones publicadas, ha estado muy presente a lo largo de todo el proceso. Debido a esto una línea de investigación podría centrarse en elevar la frecuencia de muestreo del sistema AI-IoT completo, redefiniendo el procesado de señal, datos de entrenamiento y diseño de los modelos convolucionales. Esto podría plantear la modificación de los algoritmos de cálculo de algunos parámetros para contemplar anchos de banda más elevados, lo que podría generar resultados muy interesantes.

- **Empleo de secuencias de entrada de mayor longitud:** como en el caso anterior, la dimensión de las secuencias de audio de entrada para los modelos de CNN, fue concretándose a partir de la definición de los parámetros acústicos y los diversos estudios observados. A la luz de los resultados obtenidos respecto a tiempo de cálculo, precisión y comportamiento en pruebas de campo, el empleo de secuencias de audio superiores a 1 segundo para los parámetos psicoacústicos, permitirían una apreciación distinta de sonidos continuos en el tiempo, sobretodo para los parámetros R y F. En el ámbito de los parámetros acústicos de sala, podemos afirmar que el empleo de secuencias de más de más de 3.5 s permitiría mejorar la precisión en predicción de $RT60$ para salas con tiempos de reverberación superiores a este valor. Para el resto de parámetros, permitiría una evaluación más completa, acorde a las medidas habituales de respuestas impulsivas de sala, que son más extensa, por lo que podría contribuir a mejorar la precisión, pese al incremento de procesado que representarían señales de más duración.

- **Separación de parámetros y enfoque basado en machine learning:** esta posible línea de investigación contempla 2 enfoques distintos al empelado que se complementan y han tomado forma a lo largo del desarrollo de la investigación. Si bien desde un principio se fijó la idea de predecir cada conjunto de parámetros de forma simultánea mediante un único modelo de CNN en cada caso, en el caso de los parámetros acústicos de sala sería de interés investigar no una división individual de los parámetros sino una división por naturalezas (reverberación, energéticos o de inteligibilidad), en 3 modelos que pueden concurrir en una única salida con la predicción de los 5 parámetros. Si esta es una línea de investigación interesante, plantea la posibilidad ahora sí factible del empleo de un enfoque basado en machine learning o aprendizaje automático, empleando como entrada características extraídas del audio, si bien con diferentes características para cada grupo de parámetros según su naturaleza. Este es un enfoque que en las pruebas realizadas, con las mismas características para todos los parámetros, representó un incremento de cálculo excesivo, pero que quizás podría ser asumido con la división de parámetros sugerida y dado el interés que puede suscitar el nuevo planteamiento.
- **Empleo de otros enfoques como redes LSTM:** debido a que las señales de audio de entrada son continuas en el tiempo, y las secuencias presentes guardan relación en sus características con las pasadas, sería interesante evaluar un enfoque basado en redes neuronales recurrentes, como por ejemplo una red con memoria a corto y largo plazo o *Long Short Time Memory* (LSTM), muy empleadas a la hora de afrontar problemas con señales continuas en el tiempo, tanto acústicas como de otro tipo de sensores. Estas han permitido mejorar los resultados de redes neuronales carentes de memoria en algunos casos, y no obstante, tanto el empleo de una red LSTM como de otro tipo, abre una línea de investigación muy atractiva.

Bibliografía

[1] R. Murray Schafer. *The Soundscape: Our Sonic Environment and the Tuning of the World*. Distributed to the Book Trade in the United States by American International Distribution, 1977.

[2] ISO. ISO 12913-1: 2014, Acoustics-Soundscape - Part 1: Definition and conceptual framework. Technical report, International Organization for Standardization, September 2014.

[3] ISO. ISO 1996-1: 2016, Acoustics-Description, measurement and assessment of environmental noise - Part 1: Basic quantities and assessment procedures. Technical report, International Organization for Standardization, March 2016.

[4] Environmental Noise Directive. Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the Assessment and Management of Environmental Noise, July 2002.

[5] Ronny Klæboe. Noise and health: Annoyance and interference. *Encyclopedia of Environmental Health*, pages 152–163, 12 2011.

[6] Xiangpu Gong, Benjamin Fenech, Claire Blackmore, Yingxin Chen, Georgia Rodgers, John Gulliver, and Anna L. Hansell. Association between noise annoyance and mental health outcomes: A systematic review and meta-analysis. *International Journal of Environmental Research and Public Health*, 19(5), 2022.

[7] Friederike Hammersen, Hildegard Niemann, and Jens Hoebel. Environmental noise annoyance and mental health in adults: Findings from the cross-sectional german health update (geda) study 2012. *International Journal of Environmental Research and Public Health*, 13(10), 2016.

[8] Maximo Cobos, J.J. Perez-Solano, Santiago Felici-Castell, Jaume Segura Garcia, and Juan Miguel Navarro. Cumulative-sum-based localization of sound events in low-cost wireless acoustic sensor networks. *IEEE Transactions on Audio Speech and Language Processing*, pages 1792–1802, 08 2014.

[9] Maximo Cobos, Fabio Antonacci, Anastasios Alexandridis, Athanasios Mouchtaris, and Bowon Lee. A survey of sound source localization methods in wireless acoustic sensor networks. *Wireless Communications and Mobile Computing*, 2017:1–24, 08 2017.

[10] Maximo Cobos, J.J. Perez-Solano, Oscar Belmonte Fernández, Germán Ramos, and Ana M. Torres. Simultaneous ranging and self-positioning in unsynchronized wireless acoustic sensor networks. *IEEE Transactions on Signal Processing*, 64, 11 2016.

[11] Juan Emilio Noriega-Linares, Alberto Rodriguez-Mayol, Maximo Cobos, Jaume Segura Garcia, Santiago Felici-Castell, and Juan Miguel Navarro. A wireless acoustic array

system for binaural loudness evaluation in cities. *IEEE Sensors Journal*, PP:1–1, 09 2017.

[12] J. J. Perez-Solano M. Cobos and L. T. Berger. Acoustic-based technologies for ambient assisted living. In Andrew Sirkka Sari Merilampi, editor, *Introduction to Smart eHealth and eCare Technologies*, chapter 9, pages 159–177. Taylor & Francis Group, Boca Raton, FL, USA, 2016.

[13] Raspberry Pi 4. `https://www.raspberrypi.org/products/raspberry-pi-4-model-b/`, Last accessed 21/07/2021, 2021.

[14] NVIDIA Jetson Nano. `https://www.nvidia.com/es-es/autonomous-machines/embedded-systems/jetson-nano/`, Last accessed 21/07/2021, 2021.

[15] Tinker board. `https://www.asus.com/es/Networking-IoT-Servers/AIoT-Industrial-Solution/All-series/Tinker-Board/`, Last accessed 21/07/2021, 2021.

[16] Germán Fabregat, Jose A. Belloch, José M. Badía, and Maximo Cobos. Design and implementation of acoustic source localization on a low-cost iot edge platform. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(12):3547–3551, 2020.

[17] Juan Vera-Diaz, Daniel Pizarro, and Javier Macias-Guarasa. Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates. *Sensors*, 18:3418, 10 2018.

[18] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 892–900, USA, 2016. Curran Associates Inc.

[19] Theodore Giannakopoulos, Evaggelos Spyrou, and Stavros J. Perantonis. Recognition of urban sound events using deep context-aware feature extractors and handcrafted features. In John MacIntyre, Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis, editors, *Artificial Intelligence Applications and Innovations*, pages 184–195, Cham, 2019. Springer International Publishing.

[20] Emad Grais, Mehmet Umut Sen, and Hakan Erdogan. Deep neural networks for single channel source separation. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 11 2013.

[21] Jivitesh Sharma, Ole-Christoffer Granmo, and Morten Goodwin. Environment sound classification using multiple feature channels and attention based deep convolutional neural network. INTERSPEECH 2020, October 2020.

[22] Wenjie Mu, Bo Yin, Xianqing Huang, Jiali Xu, and Zehua Du. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*, 11, 11 2021.

[23] Muqing Deng, Tingting Meng, Jiuwen Cao, Shimin Wang, Jing Zhang, and Huijie Fan. Heart sound classification based on improved mfcc features and convolutional recurrent neural networks. *Neural Networks*, 130, 06 2020.

[24] Jesus Lopez-Ballester, Adolfo Pastor-Aparicio, Jaume Segura-Garcia, Santiago Felici-Castell, and Maximo Cobos. Computation of psycho-acoustic annoyance using deep neural networks. *Applied Sciences*, 9(15):3136, Aug 2019.

[25] J. Lopez-Ballester, A. Pastor-Aparicio, S. Felici-Castell, J. Segura-Garcia, and M. Cobos. Enabling real-time computation of psycho-acoustic parameters in acoustic sensors using convolutional neural networks. *IEEE Sensors Journal*, 20(19):11429–11438, 2020.

[26] Jesus Lopez-Ballester, Jose Calero, Jaume Segura Garcia, Santiago Felici-Castell, Miguel Garcia-Pineda, and Maximo Cobos. Speech intelligibility analysis and approximation to room parameters through the internet of things. *Applied Sciences*, 11:1430, 02 2021.

[27] J. Lopez-Ballester, S. Felici-Castell, and M. Segura-Garcia, J.and Cobos. Ai-iot platform for blind estimation of room acoustic parameters based on deep neural network. *IEEE Internet of Things Journal*, 09 2022.

[28] K. Levak, M. Horvat, and H. Domitrovic. Effects of noise on humans. *2008 50th International Symposium ELMAR*, 1:333–336, 2008.

[29] Romain Sordello, Ophélie Ratel, Frédérique Flamerie De Lachapelle, Clément Leger, Alexis Dambry, and Sylvie Vanpeene. Evidence of the impact of noise pollution on biodiversity: a systematic map. *Environmental Evidence*, 9(1):20, Sep 2020.

[30] Hansjoerg P. Kunc and Rouven Schmidt. The effects of anthropogenic noise on animals: a meta-analysis. *Biology Letters*, 15(11):20190649, 2019.

[31] Natacha Aguilar Ana Tejedor. Documento técnico sobre impactos y mitigación de la contaminación acústica marina. Technical Report NIPO: 280-12-232-2, Ministerio de Agricultura, Alimentación y Medio Ambiente, 2012.

[32] Bengang Li, Shu Tao, and R.W. Dawson. Evaluation and analysis of traffic noise from the main urban roads in beijing. *Applied Acoustics*, 63(10):1137 – 1142, 2002.

[33] NoiseMap Ltd. Noise Map, Environmental Noise Mapping Software. `http://www.londonnoisemap.com`, Last accessed 16-09-2021, 2018.

[34] José Antonio Teixeira Vitienes. Servicios Smart Santander en el ámbito de la eficiencia medioambiental. `https://www.esmartcity.es/comunicaciones/servicios-smart-santander-ambito-eficiencia-medioambiental`, Last accessed 09-09-2021, 2016.

[35] Hugo Fastl and Eberhard Zwicker. *Psychoacoustics: Facts and Models*. Springer-Verlag, Berlin, Heidelberg, 2007.

[36] Brian C. J. Moore. *An Introduction to the Psychology of Hearing*. Brill, 2013.

[37] Sabine W. C. *Collected Papers on Acoustics*. Peninsula Publishing; Reprint edición, 2016.

[38] ISO. ISO 3382-1:2009. Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces. Technical report, UNE-EN, March 2016.

[39] ISO. ISO 3382-2:2008. Acoustics – Measurement of room acoustic parameters – Part 2: Reverberation time in ordinary rooms. Technical report, UNE-EN, December 2008.

[40] ISO. ISO 3382-3:2012. Acoustics – Measurement of room acoustic parameters – Part 3: Open plan offices. Technical report, UNE-EN, October 2012.

[41] ISO. ISO 226: 2003, Acoustics — Normal equal-loudness-level contours. Technical report, International Organization for Standardization, August 2003.

[42] Angelique Scharine, Kara Cave, and Tomasz Letowski. *Auditory perception and cognitive performance*, pages 391–490. U.S. Army Aeromedical Research Laboratory, 01 2009.

[43] H. Fastl. Temporal masking effects: I. broad band noise masker. *Acta Acustica united with Acustica*, 35, 08 1976.

[44] Kim Fluitt, Tomasz Letowski, and Tim Mermagen. Auditory performance in an open sound field. *The Journal of the Acoustical Society of America*, 113:2286–2286, 04 2003.

[45] Stanley A. Gelfand. *Hearing: An Introduction to Psychological and Physiological Acoustics*. CRC Press, 2017.

[46] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248, 1961.

[47] Brian Moore and Brian Glasberg. A revision of zwicker's loudness model. *Acta Acustica united with Acustica*, 82:335–345, 03 1996.

[48] ISO. UNE-EN 61260-1:2014. Electroacoustics - Octave-band and fractional-octave-band filters - Part 1: Specifications. Technical report, UNE-EN, August 2014.

[49] ISO. ISO 532-1:2017. Acoustics — Methods for calculating loudness — Part 1: Zwicker method. Technical report, UNE-EN, June 2017.

[50] DIN. DIN 45631/A1:2010-03. Calculation of loudness level and loudness from the sound spectrum - Zwicker method - Amendment 1: Calculation of the loudness of time-variant sound. Technical report, Deutsches Institut Fur Normung E.V. (German National Standard), March 2010.

[51] DIN. DIN 45692:2009-08. Measurement technique for the simulation of the auditory sensation of sharpness. Technical report, Deutsches Institut Fur Normung E.V. (German National Standard), August 2009.

[52] Roland Sottek and Julian Becker. Psychoacoustic roughness standard. *The Journal of the Acoustical Society of America*, 145:1898–1898, 03 2019.

[53] P. Daniel and R. Weber. Psychoacoustical roughness: Implementation of an optimized model. *Acta Acustica united with Acustica*, 83(1):113–123, 1997.

[54] Vincent J.P. Jourdes. Estimation of perceived roughness. Master's thesis, Industrial Engineering and Innovation Sciences Dept., Institut National Polytechnique de Grenoble, 2004.

[55] Dieter Gottlob. *Vergleich objektiver akustischer Parameter mit Ergebnissen subjektiver Untersuchungen an Konzertsälen*. Georg August Universität zu Göttingen., 1973.

[56] M. R. Schroeder, D. Gottlob, and K. F. Siebrasse. Comparative study of european concert halls: correlation of subjective preference with geometric and acoustic parameters. *The Journal of the Acoustical Society of America*, 56(4):1195–1201, 1974.

[57] T. Yamamoto and F. Suzuki. Multivariate analysis of subjective measures for sound in rooms and the physical values of room acoustics. *J. Acoust. Soc. Jpn*, 32(10):599–605, 1976.

[58] John S. Bradley and Gilbert A. Soulodre. The influence of late arriving energy on spatial impression. *The Journal of the Acoustical Society of America*, 97(4):2263–2271, 1995.

[59] J. S. Bradley and G. A. Soulodre. Objective measures of listener envelopment. *The Journal of the Acoustical Society of America*, 98(5):2590–2597, 1995.

[60] M. Barron. Late lateral energy fractions and the envelopment question in concert halls. *Applied Acoustics*, 62(2):185–202, 2001.

[61] Ando Yoichi. *Concert Hall Acoustics*. Springer-Verlag, Berlin, Heidelberg, 1985.

[62] Beranek Leo. *Concert Halls and Opera Houses*. Springer-Verlag, New York, 2004.

[63] Salvador Cerdá, Alicia Giménez, Jinson Romero, Rosa Cibrian, and J. Miralles. Room acoustical parameters: A factor analysis approach. *Applied Acoustics*, pages 97–109, 01 2009.

[64] ISO. ISO 9921: 2004, Ergonomics - Assessment of speech communication. Technical report, UNE-EN, June 2008.

[65] ANSI/ASA. Ansi/asa s3.5-1997 (r2017), american national standards institute. acoustical society of america. methods for calculation of the speech intelligibility index. Technical report, American National Standards Institute, Acoustical Society of America, June 1997.

[66] Elisa Vargas and Fausto Rodriguez. Analysis of the impact of sound diffusion in the reverberation time of an architectural space - a proposal for the characterization of diffusive surfaces using scale models. *The Journal of the Acoustical Society of America*, 123:3609, 06 2008.

[67] Gabriel Mello Silva, Alexandre Maiorino, and Stelamaris Bertoli. Study of energy acoustical parameters at audience area in a simulated multi-purpose hall with an articulated orchestra shell. In *The International Symposium on Musical and Room Acoustics*, 09 2016.

[68] Michael Vorlaender and Malte Kob. Practical aspects of mls measurements in building acoustics. *Applied Acoustics*, 52:239–258, 11 1997.

[69] Stefan Weinzierl and Michael Vorlaender. Room acoustical parameters as predictors of room acoustical impression: What do we know and what would we like to know? *Acoustics Australia*, 43:41–48, 04 2015.

[70] Michael Vorländer. Models and algorithms for computer simulations in room acoustics. In *International Seminar on Virtual Acoustics*, 2011.

[71] M. R. Schroeder. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(3):409–412, 1965.

[72] L. Gerald Marshall. An acoustics measurement program for evaluating auditoriums based on the early/late sound energy ratio. *The Journal of the Acoustical Society of America*, 96(4):2251–2261, 1994.

[73] H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *Acoustical Society of America Journal*, 67(1):318–326, January 1980.

[74] IEC. IEC 60268-16:2020-09. Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index . Technical report, International Electrotechnical Commission, September 2020.

[75] J.S. Bradley, R. Reich, and S.G. Norcross. A just noticeable difference in c50 for speech. *Applied Acoustics*, 58(2):99–108, 1999.

[76] J. Segura-Garcia, S. Felici-Castell, J. J. Perez-Solano, M. Cobos, and J. M. Navarro. Low-cost alternatives for urban noise nuisance monitoring using wireless sensor networks. *IEEE Sensors Journal*, 15(2):836–844, Feb 2015.

[77] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi. Internet of things for smart cities. *IEEE Internet of Things Journal*, 1(1):22–32, Feb 2014.

[78] J. Segura-Garcia, J. J. Perez-Solano, M. Cobos, E. Navarro, S. Felici-Castell, A. Soriano, and F. Montes. Spatial statistical analysis of urban noise data from a wasn gathered by an iot system: Application to a small city. *Applied Sciences*, 6:380, 11 2016.

[79] Jesus Lopez-Ballester, Rafael Fayos-Jordan, Jaume Segura-Garcia, Santiago Felici-Castell, Juan J. Perez-Solano, Maximo Cobos, Rosa Cibrian, and Alicia Gimenez-Perez. Análisis de inteligibilidad y aproximación a parámetros de sala mediante internet de las cosas. *Acustica 2020 - XI Congreso Ibérico de Acústica y 51º Congreso Español de Acústica - TECNIACUSTICA'20*, 2020.

[80] Jesus Lopez-Ballester, Santiago Felici Castell, Jaume Segura-Garcia, Juan José Perez-Solano, and Antonio Soriano-Asensi. Sistema iot para la monitorización de parámetros de sala e inteligibilidad del habla. *XV Jornadas de Ingenieria Telematica - JITEL2021*, 2021.

[81] Jaume Segura-García, Jesus Lopez-Ballester, Adolfo Pastor-Aparicio, Santiago Felici-Castell, Máximo Cobos-Serrano, Pérez-Solano, and García-Pineda. Visualization of nuisance information in acoustic environments using an iot system. *48th Inter-Noise*, 2019.

[82] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 1041–1044, New York, NY, USA, 2014. ACM.

[83] Adolfo Pastor-Aparicio, Jaume Segura-Garcia, Jesus Lopez-Ballester, Santiago Felici-Castell, Miguel García-Pineda, and Juan J. Pérez-Solano. Psychoacoustic annoyance implementation with wireless acoustic sensor networks for monitoring in smart cities. *IEEE Internet of Things Journal*, 7(1):128–136, 2020.

[84] A. Pastor-Aparicio, J. Lopez-Ballester, J. Segura-Garcia, S. Felici-Castell, M. Cobos, R. Fayos-Jordan, and J.J. Perez-Solano. Zwicker's annoyance model implementation in a wasn node. *48th Inter-Noise*, 2019.

[85] J. Garofolo, Lori Lamel, W. Fisher, Jonathan Fiscus, D. Pallett, N. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*, 11 1992.

[86] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[87] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[88] Y. Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35:1798–1828, 08 2013.

[89] Maximo Cobos, Jens Ahrens, Konrad Kowalczyk, and Archontis Politis. An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022, 05 2022.

[90] Hannes Gamper and Ivan J. Tashev. Blind reverberation time estimation using a convolutional neural network. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 136–140, 2018.

[91] J.M. Fields, R.G. De Jong, T. Gjestland, I.H. Flindell, R.F.S. Job, S. Kurra, P. Lercher, M. Vallet, T. Yano, R. Guski, U. Felscher-Suhr, and R. Schumer. Standardized general-purpose noise reaction questions for community noise surveys: Research and a recommendation. *Journal of Sound and Vibration*, 242(4):641–679, 2001.

[92] Edgar Tristán Hernández, Ignacio Pavón García, Juan Manuel López Navarro, Isaak Campos-Cantón, and Eleazar Samuel Kolosovas-Machuca. Evaluation of psychoacoustic annoyance and perception of noise annoyance inside university facilities. *International Journal of Acoustics and Vibration*, 23, 2018.

[93] ISO. ISO/TS 15666:2021, Acoustics — Assessment of noise annoyance by means of social and socio-acoustic surveys. Technical report, International Organization for Standardization, August 2021.

[94] ISO. ISO/TS 12913-2: 2018, Acoustics-Soundscape - Part 2: Data collection and reporting requirements. Technical report, International Organization for Standardization, August 2018.

[95] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017.

[96] Stephen G. McGovern. Fast image method for impulse response calculations of box-shaped rooms. *Applied Acoustics*, 70(1):182–189, 2009.

[97] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[98] Pablo Peso Parada, Dushyant Sharma, Jose Lainez, Daniel Barreda, Toon van Waterschoot, and Patrick A. Naylor. A single-channel non-intrusive c50 estimator correlated with speech recognition performance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):719–732, 2016.

[99] Prem Seetharaman, Gautham J. Mysore, Paris Smaragdis, and Bryan Pardo. Blind estimation of the speech transmission index for speech quality prediction. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 591–595, 2018.

[100] Nicholas Bryan. Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation. pages 1–5, 05 2020.

[101] Suradej Duangpummet, Jessada Karnjana, Waree Kongprawechnon, and Masashi Unoki. Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response. *Applied Acoustics*, 185:108372, 2022.

[102] Francesco Martellotta. The just noticeable difference of center time and clarity index in large reverberant spaces. *The Journal of the Acoustical Society of America*, 128:654–63, 08 2010.

[103] Michelle C. Vigeant, Robert D. Celmer, Chris M. Jasinski, Meghan J. Ahearn, Matthew J. Schaeffler, Clothilde B. Giacomoni, Adam P. Wells, and Caitlin I. Ormsbee. The effects of different test methods on the just noticeable difference of clarity index for music. *The Journal of the Acoustical Society of America*, 138(1):476–491, 2015.

[104] James Eaton, Nikolay Gaubitch, Alastair Moore, and Patrick Naylor. The ace challenge - corpus description and performance evaluation. 10 2015.

[105] Jaume Segura-Garcia, Adolfo Pastor-Aparicio, Jesus Lopez-Ballester, Juan J. Perez-Solano, Santiago Felici-Castell, Maximo Cobos-Serrano, Francisco Grimaldo-Moren, Miguel Arana-Burgui, Francesc Daumal-Domènech, Rosa Cibrián Ortiz de Anda, and Alicia Giménez Pérez. Descripción del paisaje sonoro de las fallas de valència (2). percepción subjetiva y parámetros psicoacústicos. *FIA 2018; XI Congreso Iberoamericano de Acústica; X Congreso Ibérico de Acústica; 49º Congreso Español de Acústica -TECNIACUSTICA'18*, 2018.

[106] Jaume Segura Garcia, Adolfo Pastor-Aparicio, Jesus Lopez-Ballester, Juan J. Perez-Solano, Santiago Felici-Castell, Máximo Cobos-Serrano, José Montoya-Belmonte, Ana Torres-Aranda, and Juan M. Navarro-Ruiz. Análisis espacio-temporal de parámetros psico-acústicos en entornos acústicos urbanos usando sistemas iot en tiempo real. *FIA 2018; XI Congreso Iberoamericano de Acústica; X Congreso Ibérico de Acústica; 49º Congreso Español de Acústica -TECNIACUSTICA'18*, 2018.

[107] Adolfo Pastor-Aparicio, Jesus Lopez-Ballester, Santiago Felici-Castell, Jaume Segura-Garcia, Rafael Fayos-Jordán, and Miguel Garcia-Pineda. Efficient implementation of an iot deployment for sound-scape monitoring. *XIV Jornadas de Ingeniería Telemática - JITEL*, 2019.

[108] Adolfo Pastor-Aparicio, Jesus Lopez-Ballester, Santiago Felici-Castell, Jaume Segura-Garcia, Rafael Fayos-Jordán, Juan José Pérez-Solano, Máximo Cobos, and Miguel Garcia-Pineda. Urbauramon: Herramientas inteligentes para la gestión y control del paisaje sonoro urbano. *XIV Jornadas de Ingeniería Telemática - JITEL*, 2019.

[109] Jaume Segura-Garcia, Santiago Felici-Castell, Jose M. Alcaraz Calero, Qi Wang, Jesus Lopez-Ballester, Rafael Fayos-Jordán, Juan J. Pérez-Solano, and Miguel Arana-Burgui. Sistema basado en tecnología 5g de monitorización psicoacústica del paisaje sonoro en smart cities con offloading computacional dinámico en el edge. *Acustica 2020 - XI Congreso Ibérico de Acústica y 51º Congreso Español de Acústica - TECNIACUSTICA'20*, 2020.

[110] Jesus Lopez-Ballester, Santiago Felici Castell, Jaume Segura-Garcia, Juan José Perez-Solano, and Antonio Soriano-Asensi. Urbauramon: Herramientas inteligentes para la gestión y monitorización acústica. *XV Jornadas de Ingenieria Telematica - JITEL2021*, 2021.

[111] Jaume Segura-Garcia, Jesus Lopez-Ballester, Santiago Felici-Castell, Juan J. Perez-Solano, Jose M. Alcaraz-Calero, Rafael Fayos-Jordan, Enrique A. Navarro-Camba, Antonio Soriano-Asensi, and Juan M. Navarro-Ruiz. Soundscape monitoring of modified psychoacoustic annoyance with next-generation edge computing and iot. *Proceedings of the 11th Euro American Conference on Telematics and Information Systems - EATIS 2022*, 2022.

Anexos

.

A. Computation of Psycho-Acoustic Annoyance Using Deep Neural Networks

applied sciences

MDPI

Article

Computation of Psycho-Acoustic Annoyance Using Deep Neural Networks

Jesus Lopez-Ballester *,†, Adolfo Pastor-Aparicio †, Jaume Segura-Garcia †, Santiago Felici-Castell † and Maximo Cobos †

Computer Science Department, Escola Tècnica Superior d'Enginyeria, Universitat de València, 46100 Burjassot, Spain
* Correspondence: jesus.lopez-ballester@uv.es
† These authors contributed equally to this work.

Received: 28 June 2019; Accepted: 1 August 2019; Published: 2 August 2019

Abstract: Psycho-acoustic parameters have been extensively used to evaluate the discomfort or pleasure produced by the sounds in our environment. In this context, wireless acoustic sensor networks (WASNs) can be an interesting solution for monitoring subjective annoyance in certain soundscapes, since they can be used to register the evolution of such parameters in time and space. Unfortunately, the calculation of the psycho-acoustic parameters involved in common annoyance models implies a significant computational cost, and makes difficult the acquisition and transmission of these parameters at the nodes. As a result, monitoring psycho-acoustic annoyance becomes an expensive and inefficient task. This paper proposes the use of a deep convolutional neural network (CNN) trained on a large urban sound dataset capable of efficiently predicting psycho-acoustic annoyance from raw audio signals continuously. We evaluate the proposed regression model and compare the resulting computation times with the ones obtained by the conventional direct calculation approach. The results confirm that the proposed model based on CNN achieves high precision in predicting psycho-acoustic annoyance, predicting annoyance values with an average quadratic error of around 3%. It also achieves a very significant reduction in processing time, which is up to 300 times faster than direct calculation, making CNN designed a clear exponent to work in IoT devices.

Keywords: convolutional neural networks; psycho-acoustic parameters; subjective annoyance; wireless acoustic sensor networks; Zwicker model

1. Introduction

In recent years, Wireless acoustic sensor networks (WASN) have received increasing attention in the field of acoustic signal processing and machine learning research. Many novel approaches for solving classical problems such as acoustic source localization [1,2], acoustic event detection and classification, or environmental noise evaluation [3] have been revisited assuming different scenarios that consider wireless acoustic nodes [4]. Additionally, in recent decades, many studies have been carried out to measure noise pollution, most of them using the acquired sound pressure level over time on the basis of a standard [5]. An example are the traffic noise control maps developed in Beijing [6] or London [7] to describe spatially the environmental noise of such cities. These measurements give us information about the weighted amplitude of the noise, i.e., "A-weighted" scale of SPL in dB(A). However, although they are somehow correlated with other spectral metrics, they provide little information about the mechanisms of action of the environmental noise on the people's mood and the subjective annoyance produced by these sounds [8]. To establish how annoying is the sound in a specific environment, other magnitudes in addition to the sound pressure level are needed. This is where the psycho-acoustic parameters, such as loudness, sharpness, roughness or fluctuation strength [9] come in. These parameters provide a better approximation of the annoyance caused by the different types of sounds perceived by

human subjects, allowing a more precise analysis of the acoustic environment. An example of the use of these parameters using an artificial neural network for classification can be shown in [10], where the psycho-acoustic parameters are used together with subjective surveys to classify the degree of annoyance or wellness of different recordings made in crowded areas within the city. The application of the soundscape description, as defined in ISO 12913-1 to a Smart City [11], involves a real-time implementation of these psycho-acoustic parameters. Traditionally, Zwicker's annoyance model [9] has been used to describe the degree of subjective nuisance produced by a noise-producing device or even in an environment by means of the determination of different psycho-acoustic parameters, namely loudness, sharpness, roughness, and fluctuation strength, to return a sound nuisance indicator. A problem arising from the use of such parameters is the significant computing time required, which presently does not allow calculation of them in real time. To lower the computing time for IoT-based systems, different approaches have been developed [8,12,13], but no one was efficient enough to develop an autonomous real-time application for a WASN.

In this paper, we propose the use of a regression model based on convolutional neural networks (CNNs) to be used as a tool to predict the total psycho-acoustic annoyance (PA) indicator from raw audio signals. To this end, we have used a training database based on the taxonomy established by ISO 12913-2 [14] for urban sounds, using the UrbanSound8K dataset [15] expanded with several recordings of our own made in different urban spaces and situations. The trained CNN-based model is shown to be around 300 times faster than the direct computation approach, providing as well significant accuracy on the test dataset.

2. Psycho-Acoustic Annoyance Model

The evaluation of psycho-acoustic annoyance over a considerably large space (e.g., a block of buildings or a city neighborhood), can be performed by means of a WASN. Due to privacy and network speed reasons, it is essential that only one general nuisance parameter is transmitted at a certain rate, so that nuisance calculations must be performed in real time at each node to optimize the system's performance. Traditionally, psycho-acoustic annoyance has been evaluated considering Zwicker's model [9], which is the chosen model used throughout this work. This section describes the parameters involved in the computation of psycho-acoustic annoyance using Zwicker's model.

2.1. Zwicker's Psycho-Acoustic Annoyance Model

Psycho-acoustic parameters allow us to quantify the degree of subjective discomfort (or pleasantness), in terms of objective metrics that certain sounds produce in people. The Zwicker's annoyance model [9] uses 4 psycho-acoustic parameters: Loudness (L), Sharpness (S), Fluctuation Strength (F) and Roughness (R) to calculate a general annoyance parameter (PA).

This model is based on the anatomy of the human ear, which behaves like a bank of filters implemented in the cochlea, the sensory organ of hearing located within the inner ear. Thus, the frequency spectrum of psycho-acoustic metrics is analyzed in terms of *critical bands* [16] with a frequency bandwidth matching the response of the mentioned auditory filters. Sound stimuli that are very close to each other in terms of frequency are combined in the human ear in the same critical band. Serializing these critical bands, we create two different frequency scales, the Bark scale and the ERB scale, called *critical band rate scales*, one measured in *Barks* and the other one measured in *Equivalent Rectangular Bandwidth* (*ERBs*) [9].

Table 1 summarizes the parameterized formulas used for calculating the psycho-acoustic metrics [9] involved in Zwicker's model. In the following, we briefly describe the parameters involved in the computation. For further details on each one, the reader is referred to [9].

- Psycho-acoustic Annoyance (PA) is a perceptual attribute that allows an objective quantification of noise annoyance from the physical characteristics of the signal, based on the mean values of L, S, R, and F.

- Loudness (L) is a perceptual measure of the effect of the energy content of sound on the ear (intensity sensation), measured in *Sones* using a linear scale. It is standardized in ISO 532B. The process used to calculate L is based on the Specific Loudness ($L'(z)$ or L contribution for the z-th critical band, measured in *Sone/Bark*. The total L is the result of accumulating all contributions across the different bands, weighted by their specific bandwidth Δz.
- Sharpness (S) is a value of sensory human perception of unpleasantness in sounds that is caused by high frequency components. It is measured in *Aures* in a linear scale.
- Roughness (R) describes the perception of the sound fluctuation even when L or $L_{eq,T}$ (i.e., the equivalent continuous sound level) remains unchanged. It analyzes the effects with different degrees of frequency modulations (around 70 Hz) in each critical band. The basic unit for R is *Asper*. For each *ERB*, $g(z)$ is an arbitrary weighting function, m is the modulation depth of each *ERB* and k is the cross-correlation between the envelopes of the *ERB* with indexes i and $i-2$.
- Fluctuation Strength (F) describes how strongly or weakly sounds fluctuate. It depends on the frequency and depth of the L fluctuations, around 4 Hz in each *ERB*. It is measured in *Vacils*.

Table 1. Numerical expressions for PA, L, S, R, and F.

$PA = L\left(1+\sqrt{\left(\frac{(S-1.75)log(L+10)}{4}\right)^2+\left(\frac{2.18(0.4F+0.6R)}{L^{0.4}}\right)^2}\right)$
$L = \sum_{z=0}^{28Bark} L'(z)\cdot\Delta z$
$S = C_S\cdot\frac{\sum_{z=0}^{28Bark} L'(z)\cdot e^{0.171\cdot z}\cdot z\cdot\Delta z}{L}$
$R = C_R\cdot\sum_{i=0.5}^{33\,ERB}(g(z_i)\cdot m_i\cdot k_{i-2}\cdot k_i)^2$
$F = C_F\cdot\sum_{i=0.5}^{33\,ERB}(g(z_i)\cdot m_i\cdot k_{i-2}\cdot k_i)^2$

The terms C_S, C_R and C_F are calibration constants.

2.2. Signal Processing and Computing Time

To calculate the nuisance model we designed four different scripts aimed at computing each of the psycho-acoustic parameters (L, S, R, and F) from an input audio signal and one to calculate the total annoyance PA from the rest of parameters. As we need to simplify the computation, the study is oriented towards an acoustic sensor network framework in which we use a single-channel audio input per node. A sampling frequency of 16 kHz is used to reduce the number of samples to be processed. Then, annoyance is analyzed over frequencies up to 8 kHz. As performed in many other studies [8,17,18] and as suggested in Zwicker's book [9], the input signal is enframed with typical window sizes of 80 to 200 ms with 50% overlap. This applies to L, S, and R, but F needs longer windows, as it takes into account modulation frequencies of about 4 Hz and performs better with window sizes in the range going from 500 ms to 1 s. In this work, we considered windows with a length of 1 s.

To perform a descriptive measurement of the required computing time of psycho-acoustic parameters and the general annoyance parameter PA, we have used three different platforms: two personal computers and a RaspberryPi-3B. The times of the personal computers were used to obtain an average of the processing time in common desktop equipment. To get an idea of the average calculation time on a normal desktop PC, we have evaluated 1000 audios on each of the 2 computers, obtaining an average of 1000 times on each. In Table 2 you can see an average of these 2 average times, in order to have a baseline of a generic PC performance calculating the general psycho-acoustic annoyance indicator, PA, and compare it with an IoT device, such as the Raspberry Pi. The Raspberry Pi platform was used to estimate the average computing time in a real node, as it is a device widely used in WASN applications [8,13,19]. The specifications are:

- PC-1: Intel(R) Core i7-7700CPU 3.6 Ghz x64; 16 GB RAM; NVIDIA GTX 1060 6 GB.

- PC-2: 2 × Intel(R) Xenon(R)Silver 2.2 Ghz x64; 96 GB RAM; NVIDIA GTX 1080 8 GB.
- Raspberry Pi 3B: CPU + GPU: Broadcom BCM2837 Cortex-A53 (ARMv8) x64 1.2 GHz; 1 GB RAM.

Table 2. PA computing time comparison in seconds.

	L	S	R	F	PA
Avg. t PC	0.0561	0.0001	0.5152	0.6429	1.2143
RPi3B t	0.0610	0.0001	0.8520	0.7423	1.6554

We used 1000 audio clips of one second duration to carry out the computing tests. These were taken randomly from our database (discussed later), measuring the computation time of each parameter in all platforms. Table 2 shows the results of the computing times. In general, *S* is the fastest as it is taken directly from *L* as described in Table 1. The calculation of *R* and *F* are the slowest ones. Looking at the results of a Raspberry Pi 3B, it can be observed that times are well above 1 s, which does not allow real-time processing. Note also that the resulting times would considerably increase for higher sampling rates. Although the use of two desktop PCs with dedicated GPUs can be confusing, in no case has the GPU been used to perform direct calculations or CNN predictions. Especially so that the conditions of the test were the same regarding the Raspberry Pi. Therefore, both the direct calculations and the predictions have been made in the CPU and only affect the characteristics of the CPU.

A more extensive study on the calculation of psycho-acoustic parameters in a WASN using Raspberry Pi is provided in [19]. Although more powerful platforms could be used for this purpose, the deployment cost would also increase significantly [20]. Thus, these results motivate the proposal of a different computational approach capable of predicting faster PA values. The following section describes our proposal based on convolutional neural networks.

3. Materials and Methods

As discussed in the previous section, an alternative tool to the direct computation of the psycho-acoustic metrics is needed to perform real-time PA evaluation within an IoT node. In recent years, deep neural networks (DNNs) have been extensively used to solve a wide range of problems related to audio signal processing, such as audio event detection [21–23], source separation [24] or source localization [25]. In this context, CNNs have been shown to be a powerful tool for many audio-related tasks, with internal layers that are able to learn optimized features capturing those signal properties that are relevant to solve the task at hand. In this work, we propose to train a CNN to solve a regression problem, where raw audio windows are provided as input, receiving as an output the predicted PA value. To this end, the PA ground-truth values used for training are obtained by direct computation, as described in Section 2 .

The data processing, the calculation of the psycho-acoustic parameters discussed above and the development of the CNN described in this section have been performed in *Matlab R2018b* and *R2019a*, making use of the functions included in the toolboxes: *DSP System Toolbox, Deep Learning Toolbox and Matlab Coder*.

3.1. Datasets

The considered audio signals used in this work belong to the UrbanSound8K database [15], consisting of 8000 tagged audios following a taxonomy similar to the one described in ISO 12913-2 [14] for urban sounds. Please note that in our case we are not especially interested in the labels, as we do not intend to recognize audio classes. All the files with a length greater than 1 s were considered and split to obtain a total of 59,000 one-second audio segments. The files were originally recorded with different sampling frequencies from 8 kHz to 48 kHz. This database was extended with 1150 s of recordings in different cities and areas, resampling all the final segments to 16 kHz. Finally, the whole database with 60,150 audio segments was divided into three datasets for training, validation,

and test. To this end, we extracted randomly 1000 audios to generate the test dataset. The remaining 59,150 signals were used for training (80%, 47,320 signals) and validation (20%, 11,830 signals). For all the audio segments, we extracted by direct computation their corresponding PA value, to use the computed values as ground-truth annoyance during the training and validation of the system. Since calibration information is missing, we assumed a standard mapping to SPL as typically performed in audio coding. Bearing in mind that PA values range from 0 to 100, proportional to unpleasant to extremely annoying sounds, it can be observed in the histogram shown in Figure 1 that most audio segments in the database have PA values in the range going from 0 to 50, so higher accuracy is expected to be achieved by system in this range.

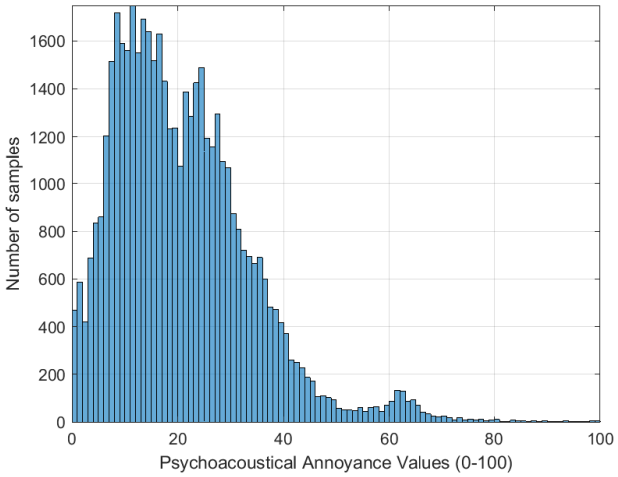


Figure 1. Histogram of PA values in the database.

3.2. Regression CNN Design and Training

The scheme summarizing the CNN architecture used in this work is shown in Figure 2. After the input layer, basically, the neural network is based on 4 *Convolutional* stages. Each of these stages consists of a block, formed by 4 layers:

- **Convolutional layer**, that applies sliding convolutional filters to the input.
- **Batch Normalization layer**, that normalizes each input channel across a mini-batch to speed up the training process.
- **Rectified Linear Unit (ReLU) layer**, which sets to zero any negative input value.
- **Max-pool layer**, that performs a down-sampling of the input by dividing the input into same size pooling regions, and computing the maximum of each region.

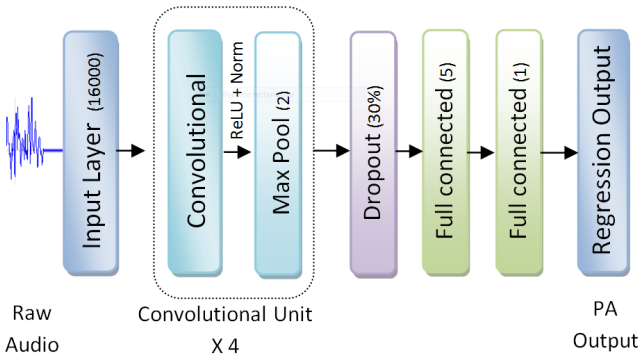


Figure 2. Regression CNN Design.

This convolutional unit is repeated 4 times in our design, as we have said and can be seen summarized in Figure 2, marked as "Convolutional Unit". In Table 3 these 4 convolutional units are shown separated by horizontal lines and you can see in detail the layers that form them. This scheme of our convolutional unit is the one that worked best when adding new convolutional layers to the overall design. This scheme allows us to progressively reduce the number of data and converge towards a very accurate PA prediction. After these 4 convolutional units a *Dropout* layer is placed with dropout probability of 0.3, followed by 2 *Fully connected* layers that reduce the output to 5 elements and then to 1, which reaches the final *Output Regression* layer that predicts the corresponding PA value. The full description of the parameters corresponding to each layer is shown in Table 3.

Table 3. CNN layer description.

Layer	Size	Filters	Stride
Input	16,000 × 1		
Convolutional S1	512 × 1	10	10
Batch Norm. S1			
ReLU. S1			
Max Pool S1	2 × 1		2
Convolutional S2	256 × 1	20	5
Batch Norm. S2			
ReLU. S2			
Max Pool S2	2 × 1		2
Convolutional S3	128 × 1	40	2
Batch Norm. S3			
ReLU. S3			
Max Pool S3	2 × 1		2
Convolutional S4	64 × 1	60	2
Batch Norm. S4			
ReLU. S4			
Max Pool S4	2 × 1		2
Convolutional S5	32 × 1	80	1
Batch Normalization S5			
ReLU			
Max Pool S5	2 × 1		2
Dropout 30%			
Fully Connected	1 × 5		
Regression Output	1 × 5		

In the column "Size" of Table 3 we describe the input dimensions of each layer, so, when we indicate that the "Input layer" has a size of 16,000 × 1 it means that its input will be a vector of 16,000 samples, which in our case refers to 1 s of audio (1 channel) sampled at 16 KHz. Each layer has an input and an output and the size of the output of one layer corresponds to the input size of the next layer. A convolutional layer applies sliding convolutional filters to the input of the layer, so that its "Size" column represents the size of each filter implemented. Taking as an example the first convolutional layer, it is formed by filters of size 512 × 1 since its entry will be a vector also (16,000 × 1). Specifically, it is formed by 10 filters of size 512 × 1, as it is indicated in the column "Filters".

Observing the Convolutional layers and the Max-Pool layers, we see that they have data in the column "Stride", this indicates us the number of elements that we move the filters before applying them again. Therefore, simplifying, we apply an A × B size filter, we get the result and we move the filter C samples before applying it again, and we repeat this operation until we reach the end of the layer input vector. Therefore, in a Max-Pool layer with size 2 × 1 and a "Stride" of 2, imagining that we move from left to right along the input data vector, we would take the largest of 2 elements, we

would move 2 samples to the right, we would take again the largest of the following 2 elements, and so on.

The training process has been carried out using the desktop computer listed as PC-2 in the Section 2.2. It is the one case of all this study in which the GPU has been used, but as we have already mentioned, only for the training of the neural network. In no case has any dedicated GPU been used to make calculations or predictions.

The training options used were as follows:

- *Solver*: SGDM
- *Momentum*: 0.9000
- *InitialLearnRate*: 1.0000×10^{-3}
- *L2Regularization*: 1.0000×10^{-4}
- *GradientThresholdMethod*: L2 Norm
- *MaxEpochs*: 70
- *MiniBatchSize*: 128
- *ValidationData*: 11,830 elements
- *ValidationFrequency*: 1107
- *Shuffle*: Every epoch

We have tried different training options to evaluate the design of our CNN, changing all the parameters mentioned above. Testing with 2, 3, and 4 convolutional stages, the options chosen are those that have yielded the best root mean square error (RMSE) results. Maintaining these options, therefore, we have tested different CNN designs with more or fewer layers and so far the configuration described in this article is the least error in prediction we get.

If we take the PA prediction, as this usually takes values in the range of 0 to 100, so observing the error made in the prediction of PA, gives us a direct idea of the percentage of error committed. Later we will discuss the error in prediction made in more detail, but we can advance by looking at Table 4 RMSE of PA prediction by using different number of stages or convolutional units. The configuration with 4 convolutional stages is a trade-off between complexity of several layers, the lowest RMSE, training and prediction times. An example of these RMSE values are shown in Table 4. It is worth mentioning that the results shown in Table 4 have been obtained by measuring in the validation dataset (11,830 signals), so the RMSE can be reduced by using the test dataset (1000 signals), as we will see later.

Table 4. Root mean square error (RMSE) with different number of convolutional stages.

Nº Conv. Stages	PA Prediction RMSE
2 stages	12.981
3 stages	4.512
4 stages	3.039

4. Evaluation and Results

After carrying out the training process of the network, the system was evaluated by using the validation dataset and through an independent test dataset (which has not been part of the training process). Considering such datasets, we measured the RMSE to evaluate the accuracy of the predictions. Also, the required time to make predictions was evaluated as well to test the speed of the CNN. The results are described in this section.

4.1. Training Results and Accuracy Test

The training process produces an RMSE value of 3.0393 by comparing the directly calculated PA of the validation dataset with the predicted PA using the trained CNN. Taking into account a

PA range from 0 to 100, this is the best value achieved after testing several CNN architectures and taking into account that PA values are considerably unbalanced (Figure 1). Similarly, we used the 1000 audio clips of the test dataset obtaining an RMSE equal to 1.7672. The improvement of the RMSE over the test dataset (1000 signals) is probably due to the fact that it is much smaller than the validation dataset (11,830 signals) and contains fewer PA values located at the extremes of the range, where most errors occur.

In Figure 3 (left) we can see a comparison between the calculated PA values (ground-truth) represented by red circles and the ones predicted by the CNN, represented by black crosses. In the right part of the same Figure 3 we can see the prediction error made by CNN. The same order of the samples is maintained so we have marked the same areas as in the left side of the figure, where the error tends to be greater due to the fact that we had fewer audio signals with PA values higher than 45 to train the CNN.

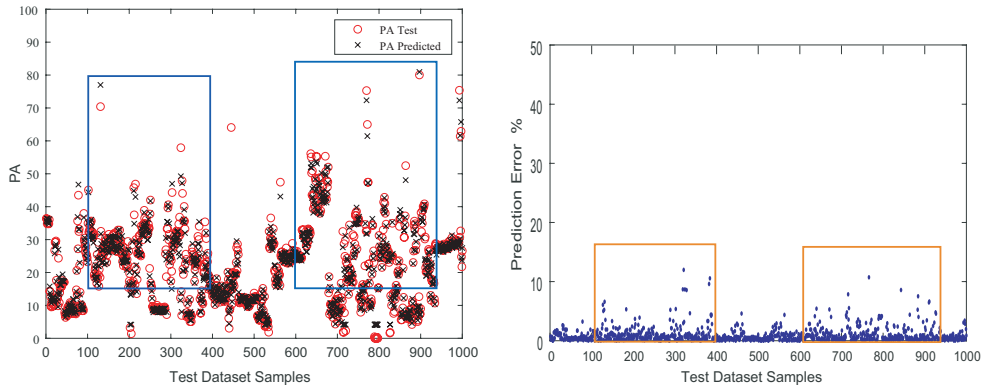


Figure 3. PA calculated vs. CNN prediction (**left**), CNN prediction error (**right**) using test dataset. Areas marked with high PA values where the highest prediction errors are located.

A scatter plot of the predicted PA values versus the calculated ground-truth values is shown in Figure 4, where we can observe more clearly the deviations of the predicted values across the PA range going from 0 to 100. In the figure we have marked with circles the values that move more than 10% away from the ideal straight line (in red). We can see that they are only 5 predictions of the 1000 made over the signals of the test dataset and that they are placed in values superior to 50 of PA, due as we have commented previously to the lack of audio samples with values of high PA in our database. As expected, the error tends to be higher for those audio segments with extremely high PA values but, overall, the CNN model provides PA predictions with very good accuracy in most cases.

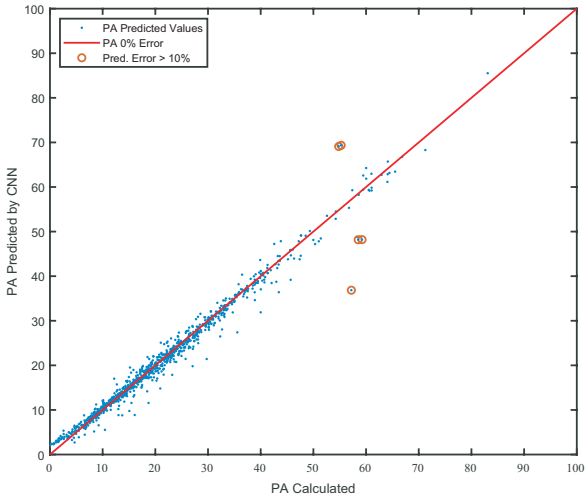


Figure 4. PA Calculated vs. PA Predicted by CNN. Predicted PA values with Error > 10% marked.

4.2. Direct Calculation vs. CNN Prediction in Terms of Computation Time

This section discusses the differences in computation time required by PA direct calculation and CNN PA prediction. To this end, we measured the time taken by the algorithm to calculate the psycho-acoustic parameters and psycho-acoustic annoyance PA, and the time taken by the neural network to predict the PA value, using the computers described in Section 2.2 in both cases. The results are shown in Figure 5.

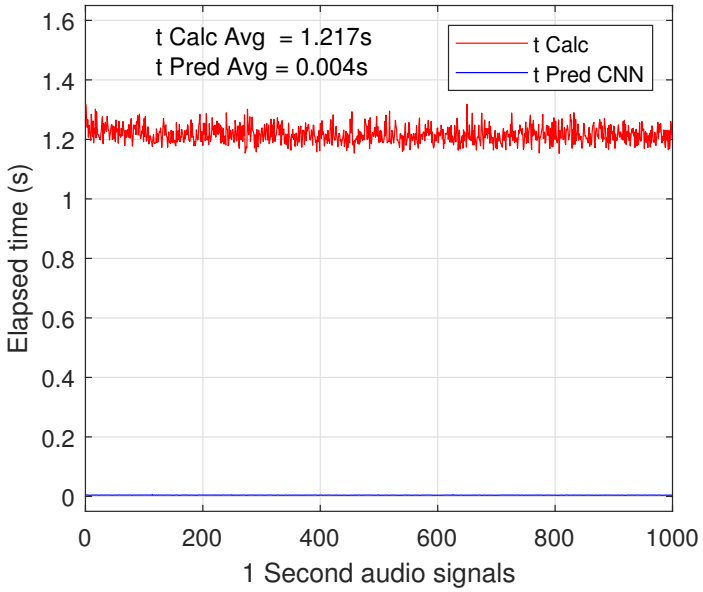


Figure 5. Elapsed time in direct calculation and elapsed time in prediction.

The time used by CNN to predict the PA value (blue) from raw audio signals is significantly smaller than the one obtained by direct calculation (red). The average time for CNN prediction is 0.004 s while for direct calculation is 1.217 s. Thus, a speedup higher than 300 is achieved by the proposed CNN-based system. It can also be observed that the times obtained from the CNN are almost constant while the times obtained by direct calculation depend on the complexity of the input sound. All these results confirm that the proposed CNN model is a good candidate for IoT-based

implementations, where the PA analysis can be comfortably performed in real time with a computation time much smaller than 1 s.

5. Conclusions

In this paper, we proposed a convolutional neural network to estimate the psycho-acoustic annoyance from raw audio signals more efficiently than by using direct calculation. The proposed CNN was shown to be very accurate in predicting the PA value of an input audio signal, quantifying discomfort according to the classical Zwicker's annoyance model. To train such model, we used a large dataset of urban sounds, using the computed annoyance over 1 s segments as ground-truth values for training and validation. Despite having a significantly unbalanced dataset, the resulting model has been confirmed to predict with a very small error the PA values in the range going from 0 to 50 PA units, and moderate errors in those audio segments presenting extremely high PA values. On the other hand, we also compared the computing time required by the proposed model and the one obtained by means of direct calculation, obtaining a speedup higher than 300. We have demonstrated the effectiveness of using deep neural networks to estimate PA in IoT devices with limited resources, performing computations in the same node and implementing a smart WASN more easily and efficiently.

Author Contributions: The authors contributed as follows: conceptualization, J.Lopez.; methodology, J.Lopez and A.Pastor.; software, J.Lopez and A.Pastor.; validation, J.Lopez, A.Pastor and J.Segura.; formal analysis, J.Lopez and M.Cobos; investigation, J.Lopez.; resources, S.Felici, M.Cobos and J.Segura.; data curation, J.Lopez; writing–original draft preparation, J.Lopez.; writing–review and editing, J.Lopez, J.Segura and S.Felici; supervision, M.Cobos, J.Segura and S.Felici; project administration, J.Segura, S.Felici and M.Cobos; funding acquisition, J.Segura, S.Felici and M.Cobos.

Funding: This work has been funded by the Spanish Ministry of Economy and Competitiveness and co-funded with European Regional Development Fund (FEDER) under the project grants with reference BIA2016-76957-C3-1-R and RTI2018-097045-B-C21.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cobos, M.; Perez-Solano, J.; Felici-Castell, S.; Segura Garcia, J.; Navarro, J.M. Cumulative-Sum-Based Localization of Sound Events in Low-Cost Wireless Acoustic Sensor Networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1792–1802. [CrossRef]
2. Cobos, M.; Antonacci, F.; Alexandridis, A.; Mouchtaris, A.; Lee, B. A Survey of Sound Source Localization Methods in Wireless Acoustic Sensor Networks. *Wirel. Commun. Mob. Comput.* **2017**, *2017*, 3956282. [CrossRef]
3. Noriega-Linares, J.E.; Rodriguez-Mayol, A.; Cobos, M.; Segura Garcia, J.; Felici-Castell, S.; Navarro, J.M. A Wireless Acoustic Array System for Binaural Loudness Evaluation in Cities. *IEEE Sens. J.* **2017**, *17*, 7043–7052. [CrossRef]
4. Cobos, M.; Perez-Solano, J.J.; Berger, L.T. Acoustic-based technologies for ambient assisted living. In *Introduction to Smart eHealth and eCare Technologies*; Sari Merilampi, A.S., Ed.; Taylor & Francis Group: Boca Raton, FL, USA, 2016; Chapter 9, pp. 159–177.
5. International Organization for Standardization (ISO). *ISO 1996-1: 2016, Acoustics-Description, Measurement and Assessment of Environmental Noise—Part 1: Basic Quantities and Assessment Procedures*; Technical Report; ISO: Geneva, Switzerland, 2016.
6. Li, B.; Tao, S.; Dawson, R. Evaluation and analysis of traffic noise from the main urban roads in Beijing. *Appl. Acoust.* **2002**, *63*, 1137–1142. [CrossRef]
7. NoiseMap Ltd. Noise Map, Environmental Noise Mapping Software. 2018. Available online: http://www.londonnoisemap.com (accessed on 19 June 2019).
8. Segura-Garcia, J.; Felici-Castell, S.; Perez-Solano, J.J.; Cobos, M.; Navarro, J.M. Low-Cost Alternatives for Urban Noise Nuisance Monitoring Using Wireless Sensor Networks. *IEEE Sens. J.* **2015**, *15*, 836–844. [CrossRef]

9. Fastl, H.; Zwicker, E. *Psychoacoustics: Facts and Models*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 327–329.
10. Maffei, L.; Masullo, M.; Toma, R.A.; Ciaburro, G.; Firat, H.B. Awaking the awareness of the movida noise on residents: Measurements, experiments and modelling. In Proceedings of the 48th Inter-Noise, Madrid, Spain, 16–19 June 2019.
11. International Organization for Standardization (ISO). *ISO 12913-1: 2014, Acoustics-Soundscape—Part 1: Definition and Conceptual Framework*; Technical Report; ISO: Geneva, Switzerland, 2014.
12. Zanella, A.; Bui, N.; Castellani, A.; Vangelista, L.; Zorzi, M. Internet of Things for Smart Cities. *IEEE Internet Things J.* **2014**, *1*, 22–32. [CrossRef]
13. Segura-Garcia, J.; Perez-Solano, J.J.; Cobos, M.; Navarro, E.; Felici-Castell, S.; Soriano, A.; Montes, F. Spatial Statistical Analysis of Urban Noise Data from a WASN Gathered by an IoT System: Application to a Small City. *Appl. Sci.* **2016**, *6*, 380. [CrossRef]
14. International Organization for Standardization (ISO). *ISO/TS 12913-2: 2018, Acoustics-Soundscape—Part 2: Data Collection and Reporting Requirements*; Technical Report; ISO: Geneva, Switzerland, 2018.
15. Salamon, J.; Jacoby, C.; Bello, J.P. A Dataset and Taxonomy for Urban Sound Research. In Proceedings of the 22nd ACM International Conference on Multimedia (MM '14), Orlando, FL, USA, 3–7 November 2014; ACM: New York, NY, USA, 2014; pp. 1041–1044. [CrossRef]
16. Gelfand, S.A. *Hearing: An Introduction to Psychological and Physiological Acoustics*; CRC Press: Boca-Raton, FL, USA, 2017.
17. Terhardt, E.; Stoll, G.; Seewann, M. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *J. Acoust. Soc. Am.* **1982**, *71*, 679–688. [CrossRef]
18. Lingsong, H.; Crocker, M.J.; Ran, Z. FFT based complex critical band filter bank and time-varying loudness, fluctuation strength and roughness. In Proceedings of the International Congress on Sound and Vibration 2007 (ICSV14), Cairns, Australia, 9–12 July 2007.
19. Pastor-Aparicio, A.; Lopez-Ballester, J.; Segura-Garcia, J.; Felici-Castell, S.; Cobos, M.; Fayos-Jordan, R.; Perez-Solano, J. Real time implementation for psycho-acoustic annoyance monitoring on wireless acoustic sensor networks. In Proceedings of the 48th Inter-Noise, Madrid, Spain, 16–19 June 2019.
20. Belloch, J.A.; Badía, J.M.; Igual, F.D.; Cobos, M. Practical Considerations for Acoustic Source Localization in the IoT Era: Platforms, Energy Efficiency and Performance. *IEEE Internet Things J.* **2019**, *6*, 5068–5079. [CrossRef]
21. Aytar, Y.; Vondrick, C.; Torralba, A. SoundNet: Learning Sound Representations from Unlabeled Video. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16), Barcelona, Spain, 5–10 December 2016; Curran Associates Inc.: Red Hook, NY, USA, 2016; pp. 892–900.
22. Giannakopoulos, T.; Perantonis, S. Recognition of Urban Sound Events using Deep Context-Aware Feature Extractors and Handcrafted Features. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Rhodes, Greece, 25–27 May 2018.
23. Martin-Morato, I.; Mesaros, A.; Heittola, T.; Virtanen, T.; Cobos, M.; J. Ferri, F. Sound Event Envelope Estimation in Polyphonic Mixtures. In Proceedings of the 2019 IEEE International Conference on Acoustics (ICASSP 2019), Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 935–939. [CrossRef]
24. Grais, E.; Umut Sen, M.; Erdogan, H. Deep neural networks for single channel source separation. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2013; doi:10.1109/ICASSP.2014.6854299. [CrossRef]
25. Vera-Diaz, J.; Pizarro, D.; Macias-Guarasa, J. Towards End-to-End Acoustic Localization Using Deep Learning: From Audio Signals to Source Position Coordinates. *Sensors* **2018**, *18*, 3418. [CrossRef] [PubMed]

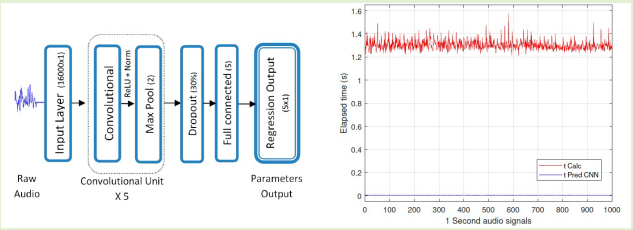
.

B. Enabling Real-Time Computation of Psycho-Acoustic Parameters in Acoustic Sensors Using Convolutional Neural Networks

Enabling Real-Time Computation of Psycho-Acoustic Parameters in Acoustic Sensors Using Convolutional Neural Networks

Jesus Lopez-Ballester, Adolfo Pastor-Aparicio, Santiago Felici-Castell, Jaume Segura-Garcia, and Maximo Cobos, *Senior Member, IEEE*

***Abstract*—Sensor networks have become an extremely useful tool for monitoring and analysing many aspects of our daily lives. Noise pollution levels are very important today, especially in cities where the number of inhabitants and disturbing sounds are constantly increasing. Psycho-acoustic parameters are a fundamental tool for assessing the degree of discomfort produced by different sounds and, combined with wireless acoustic sensor networks (WASNs), could enable, for example, the efficient implementation of acoustic discomfort maps within smart cities. However, the continuous monitoring of psycho-acoustic parameters to create time-dependent discomfort maps requires a high computational demand that prevents real-time computations within the nodes. Moreover, sending audio streams outside of the WASN for their further computation, would require extra communication and computational efforts without warranting a real-time monitoring, with the added problem of violating some privacy laws. As a result, most existing systems for nuisance assessment are usually based on less accurate indicators that require lower computational cost. In this paper, we describe the design and analysis of a deep convolutional neural network (CNN) trained with a big dataset of typical sounds occurrying in a city. The CNN allows to predict the psycho-acoustic parameters considered by the well-known Zwicker's psycho-acoustic nuisance model with great accuracy, directly from the raw recorded audio signal. The proposed CNN-based system has been tested on both desktop computers and typical WASN devices (such as Raspberry Pi), achieving very fast calculation times that allow real-time operation and a continuous monitoring of psycho-acoustic parameters.**



***Index Terms*— Annoyance, neural networks, psycho-acoustic parameters, wireless acoustic sensor networks, Zwicker annoyance model.**

I. Introduction

ACOUSTIC contamination or noise pollution is a major problem that affects the quality and health of our lives. Different rules and standards have been issued to cope with in the last decades, such as ISO 1996-1 [1], Environmental Noise Directive (END) 2002/49/EC [2] and ISO 12913:1 [3]. Behind of them, there is a great concern in urban environments for the quality of life of inhabitants in terms of subjective nuisance or annoyance, that have been demonstrated it has a big influence on the health of their citizens. Some of these standards focus on traditional Sound Pressure Levels (SPLs), or its variants (such as A-weighting picking up the frequency range around 3-6 kHz to which the human ear is more sensitive) [1], [2], while others focus on subjective nuisance or Psycho-Acoustic Annoyance (PA), that is a new trend in the last decade, closely related to the analysis of soundscapes. It is worth mentioning that PA has been recently regulated by ISO 12913:2 (Soundscape) [4].

To keep pace with such rules and standards in urban environments, it is usually required a dense and distributed acoustic sensor network. Traditionally, Wireless Acoustic Sensor Networks (WASNs) based on low cost sensing nodes have been used to address these issues due to their flexibility [5]. In the literature, WASNs have been widely used in different scenarios, for instance to build a map urban noise levels [6]

Manuscript received February 13, 2020; accepted May 17, 2020. Date of publication May 19, 2020; date of current version September 3, 2020. This work was supported in part by the Spain Government under Grant BIA2016-76957-C3-1-R and Grant RTI2018-097045-B-C21 and in part by the University of Valencia under Grant UV-INV_EPDI19-995284. The associate editor coordinating the review of this article and approving it for publication was Prof. Danilo Demarchi. *(Corresponding author: Jaume Segura-Garcia.)*

The authors are with the Department of Computer Science, ETSE, Universitat de València, 46100 Valencia, Spain (e-mail: jesus.lopez-ballester@uv.es; apasa3@alumni.uv.es; santiago.felici@uv.es; jaume.segura@uv.es; maximo.cobos@uv.es).

Digital Object Identifier 10.1109/JSEN.2020.2995779

and road traffic noise levels [7], [8], to identify animal sounds [9], [10] or other types of events or even to locate these acoustic events [11]. Usually in these scenarios, the measurements are mainly based on equivalent SPL (henceforth N_{eq}). However, these measurements are not enough in terms of annoyance assessment, due to the fact that, for instance, similar N_{eq} values can lead to different feelings of the noise perceived by different people according to its frequency characteristics. Thus, a simple N_{eq} value fails to provide enough information related to the perceived annoyance and its psycho-acoustic properties [12]. In order to build a soundscape description and to define metrics based on the human hearing system, different studies and techniques have been carried out based on the estimation of the psycho-acoustic annoyance (PA), which follows the well-known and extensively used Zwicker model [13]. In this model, PA is based on the estimation of Loudness (N), Sharpness (S), Roughness (R) and Fluctuation strength (F).

The main problem to continuously analyze a sound landscape as defined in ISO 12913-1 in a Smart City [3] is that the above parameters must be computed in real time. However, the computation of the whole set of acoustic parameters (N, S, R, F and PA) in a traditional WASNs is not an easy task, since these networks usually lack the required computational power to achieve a near real-time calculation of an accurate soundscape profiling. Nonetheless, different initiatives have been proposed to determine these parameters, such as [6], [12], [14], but non of them has been efficient enough to develop an autonomous real-time application for a WASN fulfilling the real time requirement. We must bear in mind that usually these networks are set up considering low-cost Small Board Computers (SBCs) such as Raspberry Pi [15].

Recently, the authors proposed the use of deep neural networks for predicting PA from raw audio signals, showing very promising results [16]. However, a more detailed psycho-acoustic description of the input sound considering the whole set of parameters is needed to enable an in-depth and complete analysis beyond PA, which motivates the present work. In this article, we propose an extended approach following a multi-output regression-based model using a Convolutional Neural Network (CNN). Instead of predicting only PA, the presented model considers all the aforementioned psycho-acoustic parameters (N, S, R, F and PA) and can be implemented in real-time within a conventional WASN node such as Raspberry Pi. For training purposes, we used a big training audio collection of urban sounds given by the UrbanSound8K dataset [17], which follows the requirements and criteria defined by ISO 12913-2 [4] for urban environments. This dataset is considered to be representative of the sounds present in urban soundscapes, allowing the proposed system to learn from data matching the acoustical properties of the sounds expected in this application scenario. In addition, to improve the accuracy and validate the performance over unseen examples, we have added to this database other audio samples as well as our own recordings from different urban spaces and situations. The trained CNN shows a significant speedup improvement, around 250 times faster than the direct computation of these parameters on the same nodes, with very small prediction errors. Therefore, the trained model can be successfully implemented on these nodes, allowing for real-time operation and great accuracy on these WASNs, with intended application scenarios ranging from smart cities and buildings to general public spaces (e.g. libraries, schools, museums, etc.).

The main contributions of this paper are as follows. First, we propose an end-to-end CNN-based system that allows to predict a set of meaningful psycho-acoustic parameters from raw audio signals with reduced computational resources. Such a system is shown to be up to 250 times faster than a direct calculation approach on a desktop PC, and more 58 times for an SBC (Raspberry Pi 3), enabling low-cost acoustic monitoring solutions. Second, we evaluate the trade-off between prediction accuracy and architecture complexity, showing how the number of convolutional blocks or stages in a CNN affects the final model performance for this task. Finally, we provide performance measurements for different computation alternatives on different platforms, assessing the suitability of CNN and direct computation in practical WASN implementations.

This paper is structured into five sections. Section II describes the Zwicker psycho-acoustic model providing the main framework used throughout this paper. Section III presents the background and design considerations related to the use of CNNs in our proposed system. Section IV describes the experimental evaluation and the results of the proposed system, both in terms of accuracy and computing time. Finally, the conclusions of this work are summarized in Section V.

II. Background

The deployment of WASNs enables the evaluation of psycho-acoustic nuisance in large environments (e.g., a building or a city neighbourhood). For the calculation of psycho-acoustic parameters it requires the recording and processing of audio data. Nevertheless, due to privacy reasons, sending out raw audio is forbidden as well as its transfer many times exceeds the network bandwidth. In this scenario, it is necessary that the nodes itself perform the calculation of the annoyance locally in real time, sending out periodically only a set of psycho-acoustic parameters as a result. The psycho-acoustic annoyance model used in this work is the Zwicker model [13]. The parameters making up this model are assumed to provide a more complete description of the different psycho-acoustic effects leading to the evaluation of the perceived annoyance. As described throughout the next sections, the aim of this work is to accurately predict such parameters from raw audio data with reduced computational resources. Next, the defined psycho-acoustic parameters involved in the calculation of psycho-acoustic annoyance using the Zwicker model are described.

A. Zwicker's Psycho-Acoustic Annoyance Model

Psycho-acoustic parameters are defined as objective metrics that quantify the different human sensations generated by a sound. In order to obtain a general annoyance parameter PA, the Zwicker model is based on 4 psycho-acoustic parameters: N, S, R and F. These psycho-acoustic parameters are

designed to simulate the effects produced by the human ear within the cochlea, which performs a frequency analysis of the incoming sound that behaves like a bank of filters whose bandwidth is frequency-dependent. Such filters group those sound stimuli, having frequency content close to the same critical band. For this reason, all psycho-acoustic parameters are analysed using *Critical Bands* [18], whose bandwidth is equivalent to that generated by the auditory filter. Two psycho-acoustic scales, the Bark scale and the ERB scale (both known as *Critical Band Rate Scales*), model this behaviour, one measured in *Barks* and the other one in *Equivalent Rectangular Bandwidth* (*ERBs*) [13]. Next, the parameters involved in the PA calculation are briefly described. For further details on each one, the reader is referred to [13].

1) Loudness: Loudness (N) measures the human sensation of sound intensity. When measured in *phones* using a logarithmic scale, it is referred to Loudness level and it is usually symbolized as L. It can be also measured in *sones* using the linear scale that best fits human perception, representing it as N. In our case, to implement our calculation algorithm, we have chosen this last scale, as it is more closely related to the actual human perception mechanisms. The calculation method is standardised by ISO 532B. The most common way to calculate it is to add the Specific Loudness ($N'(z)$ or N contribution for the z-th critical band, measured in *Sone/Bark*) weighted by the bandwidth belonging to each critical band:

$$N = \sum_{z=0}^{28Bark} N'(z)\Delta_z. \tag{1}$$

2) Sharpness: Sharpness (S) measures the human perception related to the amount of high frequency components making up a sound. Thus, sounds having a significant amount of high-frequency content will be considered to be sharper. Its unit is the *Acum* and can be calculated by the following the equation:

$$S = C_S \frac{\sum_{z=0}^{28Bark} N'(z)e^{0.171z} \cdot z\Delta_z}{N}. \tag{2}$$

3) Roughness: Roughness (R) is the effect that quantifies the subjective perception of fast amplitude modulation of a sound (15-300 Hz). This effect occurs when a sound fluctuates around 70Hz and is calculated for each critical band. Its unit is the *Asper*. For each *ERB*, $g(z)$ is an arbitrary weighting function, m is the modulation depth of each *ERB* and k is the cross-correlation between the envelopes of the *ERB* with indexes i and $i-2$. It is calculated as:

$$R = C_R \sum_{i=0.5}^{33\,ERB} \left(g(z_i) \cdot m_i \cdot k_{i-2} \cdot k_i\right)^2. \tag{3}$$

4) Fluctuation Strength: Fluctuation Strength (F) measures the subjective perception of slow modulations contained in a sound. This effect occurs when a sound fluctuates around 4Hz and disappears after 20Hz. Its unit is the *Vacil*. It is calculated as:

$$F = C_F \sum_{i=0.5}^{33\,ERB} \left(g(z_i) \cdot m_i \cdot k_{i-2} \cdot k_i\right)^2. \tag{4}$$

5) Psycho-Acoustic Annoyance: Psycho-acoustic Annoyance (PA) quantitatively describes the level of sound annoyance given the physical characteristics of a signal, based on the weighting of the calculated values N, S, R and F. Once the above parameters are obtained, we can calculate PA as follows:

$$PA = N\left(1 + \sqrt{{w_S}^2 + {w_{FR}}^2}\right), \tag{5a}$$

$$w_S = 0.25 \cdot (S - 1.75)\log(N + 10), \tag{5b}$$

$$w_{FR} = \frac{2.18 \cdot (0.4F + 0.6R)}{N^{0.4}}. \tag{5c}$$

The calculation of S, R and F includes 3 calibration constants, included in Eqs. (2), (3) and (4) as C_S, C_R and C_F.

B. Direct Calculation Implementation Issues

As seen in Eq. 5a, to calculate the total annoyance indicator of the Zwicker model (PA), we need first to calculate the value of the psycho-acoustic parameters. In this case, a script receiving as input the values of N, S, R and F has been implemented, returning the value of PA. Similarly, we have implemented 4 independent scripts that calculate N, S, R, and F from an input single-channel raw audio signal, as would correspond to the signal captured by a single-microphone node of a WASN. A sampling frequency of 16 KHz has been considered, which provides a good trade off between the number of samples to process and the frequency range considered by the extracted parameters.

In many studies [12], [19], [20], as suggested in Zwicker's book [13], the audio input signals are enframed considering windows of 80 to 200 ms with an overlap of up to 50%. This is valid for N or S, but insufficient for R or F, where the evaluation of very low frequency modulations (of the order of 4Hz) requires longer audio windows. As it will be later described in detail when presenting the deep neural network approach, in our case we have chosen 1 s audio windows without overlap.

Since the main objective of our work is to relax the computational needs and calculation times employed to obtain the psycho-acoustic parameters, we first evaluate the time taken by common devices to perform this task. We considered 2 desktop PCs with different characteristics as a baseline, and a Raspberry PI-3B+ as an example IoT platform of a WASN device [6], [12], [21]. The equipment specifications are shown in Table I. To measure the average calculation time required by each parameter, we considered 1000 random audio files of 1 s duration. The resulting average calculation time for each parameter is shown in Table II. While S and N are faster to compute, R and F are considerably slower due to the filtering operations needed. The PA time is a good indicator that summarizes the performance of each device, as it needs the pre-computation of all the other parameters. It must be stressed that all of them require times longer than the duration of the input signal (1 s), preventing real-time operation. Although platforms with higher computing power could be used, the cost of such deployment would

TABLE I
TEST DEVICES SPECS

	CPU	Cores	CPU Speed	RAM
PC 1	Intel i7-7700 x64	4	3.6 GHz	16 GB
PC 2	2× Intel Xeon x64	2×10	2.2 GHz	96 GB
RPi3B+	ARM Cortex A-53 x64	4	1.4 GHz	1 GB

TABLE II
PSYCHO-ACOUSTIC PARAMETERS COMPUTING TIME COMPARISON IN SECONDS

	N	S	R	F	PA
PC 1	0.0561	0.0001	0.5152	0.6429	1.2143
PC 2	0.0579	0.0001	0.5287	0.6511	1.2379
RPi3B+	0.0610	0.0001	0.8520	0.7423	1.6554

increase too much [22], moving away from the concept of a traditional WASN node. A detailed discussion of the problems related to the computation of psycho-acoustic parameters in WASN nodes can be found in [21], motivating the CNN-based approach presented in the following section.

III. CONVOLUTIONAL NEURAL NETWORK

As already described, an alternative way to perform the calculation of the psycho-acoustic parameters in real time is needed. Recently, deep neural networks have attracted the attention of the acoustic signal processing research community, for addressing tasks such as source localization [23], event detection [24], [25] or source separation [26]. In this context, CNNs have been successfully applied to solve these problems thanks to their ability to learn optimized feature representations. In a very schematic way, a neural network is a set of layers formed by interconnected "neurons" which, after a training procedure, provide us with certain information on the output for a given input. In a convolutional neural network, in certain layers the parameters in the neurons are filters that carry out convolutions, so that they activate towards certain inputs. During the training process, the weights of the filters in the convolutional layers are adjusted so that the desired parameters at the output are obtained from the input signals. It is very common to find CNNs dedicated to data classification, for example, classifying the places that appear in videos, from images or from their audio stream [24]. However, the output can not only be a class probability, but it can take the form of one or more numerical values depending on the input present in the CNN. In this case a regression layer with linear activation is used at the end of the CNN to make a prediction of the desired output value(s). This is the case here because we want to predict the values of the psycho-acoustic parameters (L, S, R, F) and the psycho-acoustic annoyance (PA) from an input audio signal. In this article, we propose the use of a CNN to solve our problem using a regression-based approach, providing raw audio signals as input and obtaining at the output the predicted values of the set of psycho-acoustic parameters: N, S, R, F in addition to the general annoyance indicator PA.

The use of an end-to-end CNN architecture is motivated by their ability to learn filters focusing on different frequency bands of the input [24], which may perform tasks similar to the analysis filter banks used by the direct computation algorithms. To this end, ground-truth parameter values used for training are obtained by direct computation, as described in Section II.

A. Datasets

The considered audio signals used in this work belong to the UrbanSound8K database [17], consisting of 8000 tagged audios following a taxonomy similar to the one described in ISO 12913-2 [4] for urban sounds. Note that in our case, we are not especially interested in the labels, as we do not intend to recognize audio classes. All the files having a length greater than 1 s were considered and splitted to obtain a total of 59000 one-second audio segments. The files were originally recorded with different sampling frequencies from 8 kHz to 48 kHz. This database was extended with 1150 seconds of recordings from different cities and areas, resampling all the final segments to 16 kHz. Finally, the whole database with 60150 audio segments was divided into three datasets for training, validation and test. To this end, we extracted randomly 1000 audios to generate the test dataset. The remaining 59150 signals were used for training (80%, 47320 signals) and validation (20%, 11830 signals). For all the audio segments, we extracted by direct computation their corresponding N, S, R, F and PA values, in order to use the computed values as ground-truth annoyance during the training and validation of the CNN. Since calibration information is missing in the original recordings, we assumed a standard mapping to SPL as typically performed in audio coding.

In Figure 1, we can see the histograms of the calculated parameters for all the audio segments of our database. Analyzing the histograms, we see that N is the parameter that mainly influences the general psycho-acoustic annoyance PA, as we can deduce from Eq.(5a). It can be observed that most audios in the database show low R and F values due to the difficulty to find modulated sounds in audible frequencies in normal city environments, which translates into a fairly small value range for such parameters. This contributes to unbalance our training dataset, because when calculating the PA of real-world city sounds, it is not usual to find extreme values presenting too little or too high annoyance. Bearing in mind that PA values range from 0 to 100, proportional to slightly unpleasant to extremely annoying sounds, it can be seen from the lowest histogram in Figure 1 that most of the audio segments in the database have PA values in the range from 0 to 50. Thus, it is expected that the system trained with these data will achieve greater accuracy in this range. We also expect less accuracy in predicting R and F, since we have a limited range of values in the database.

B. CNN Design and Training

The proposed architecture is based on SoundNet [24], which is a well-known CNN-based architecture for audio classification. Note that, as in many other deep learning approaches (e.g. AlexNet), the size of the filters decreases from the input to the output, while, conversely, their number is increased. In this context, the selection of the hyperparameters in our system was selected by testing different architectures, and we fine-tuned them until achieving reasonable performance.

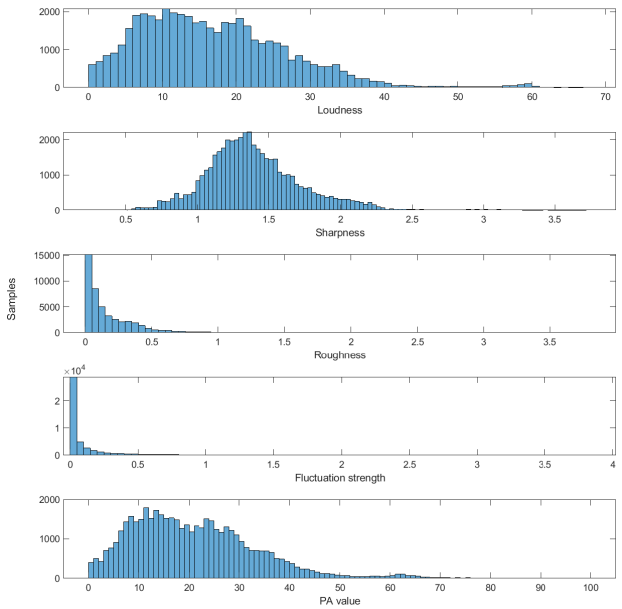


Fig. 1. Histogram of computed *N*, *S*, *R*, *F* and *PA* values in our training dataset.

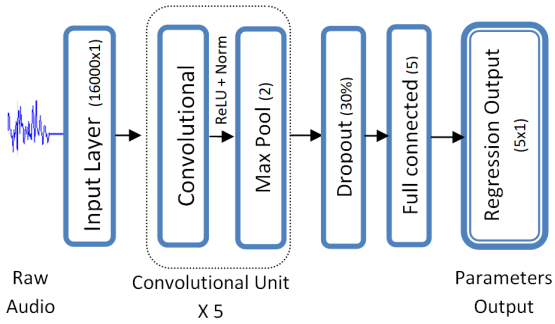


Fig. 2. Proposed CNN architecture.

The scheme summarizing the CNN architecture used in this work is shown in Figure 2. Basically, the neural network is based on 5 *Convolutional* stages (from S1, stage 1, till S5, stage 5). Each of these stages consists of a temporal convolution block, followed by a *Batch Normalization* layer that normalizes each input channel across a mini-batch to speed up the training process. After this layer, it follows a *Rectified Linear Unit (ReLU)* activation layer, which sets to zero any negative input value. Finally, each convolutional unit ends with a *Max-pool* layer that keeps the maximum of the indicated elements, implementing a downsampling of the input. This convolutional unit is repeated five times. A *Dropout* layer with dropout probability of 0.3 is placed after the convolutional units to prevent overfitting, followed by a *Fully connected* layer that reduces the output to 5 elements, reaching the final *Output Regression* activation layer (linear) that predicts the corresponding N, S, R, F and PA values. The full description of the parameters corresponding to each layer is shown in Table III. It must be noticed that the selected network depth is considerably smaller than the one used by most state-of-the-art CNNs for audio analysis tasks, such as acoustic event detection and audio scene classification [27], [28], because the parameter estimation tasks can be considered an easier problem.

TABLE III
CNN LAYERS DESCRIPTION

Layer	Size	Filters	Stride
Input	16000×1		
Convolutional S1	512×1	10	10
Batch Norm. S1			
ReLU. S1			
Max Pool S1	2×1		2
Convolutional S2	256×1	20	5
Batch Norm. S2			
ReLU. S2			
Max Pool S2	2×1		2
Convolutional S3	128×1	40	2
Batch Norm. S3			
ReLU. S3			
Max Pool S3	2×1		2
Convolutional S4	64×1	60	2
Batch Norm. S4			
ReLU. S4			
Max Pool S4	2×1		2
Convolutional S5	32×1	80	1
Batch Normalization S5			
ReLU			
Max Pool S5	2×1		2
Dropout 30%			
Fully Connected	1×5		
Regression Output	1×5		

TABLE IV
RMSE WITH DIFERENT NUMBER OF CONVOLUTIONAL STAGES

Nº Conv. Stages	PA prediction RMSE
2 *stages*	12.9810
3 *stages*	4.5120
4 *stages*	3.0390
5 *stages*	2.5306
6 *stages*	3.8480

The training process has been carried out with an SGDM solver with a momentum of 0.9. We started with a learning rate of 10^{-3}, using an L2 regularization parameter of 10^{-4}. The network was training for a maximum of 80 epochs, shuffling each epoch with a minibatch size of 128 examples.

To select an appropriate number of convolutional layers, we performed several experiments changing the number of stages. The results for the PA prediction error can be observed in Table IV, where it can be observed that the configuration based on 5 convolutional stages, is the one that achieves the lowest *Root Mean Square Error* (RMSE).

IV. RESULTS

This section analyzes the performance, error and computation time of our proposed deep learning solution based on the previously described CNN architecture. First, we visualize the first layer of the network by looking at the learned weights.

A. Layer Weights in Convolutional S1

We provide below an analysis of the filters of the first convolutional layer in the network, which consists of 10 filters of

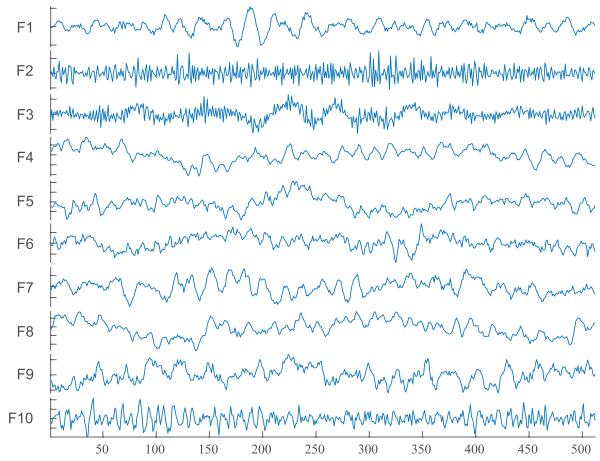


Fig. 3. Convolutional S1 layer weights.

size 512×1. The plot of the filters (which are unidimensional) are shown in Figure 3.

It should be emphasized that the first convolutional layer is where the first feature extraction process takes place from the raw audio input. It can be observed that the learnt filters reflect diverse spectro-temporal characteristics, such as different frequency ripplings or temporal modulations. Note, for example, that filter 3 (column 3) shows a low-frequency temporal modulation over a high frequency, while filters tuned to lower frequencies are observed in columns 1 and 4. This behavior roughly resembles some of the characteristics related to the algorithms involved in the direct calculation of the psycho-acoustic parameters: e.g. in Loudness or Sharpness, the SPL levels of the acoustic input are evaluated as a function of perceptual frequency bands, while Fluctuation Strength and Roughness may better reflect the presence of low-frequency modulations.

B. Performance

Once the training process was finished, we tested the CNN with the test dataset to predict the 4 psycho-acoustic parameters (N, S, R and F) and the psycho-acoustic annoyance (PA), then we measured the RMSE made in the prediction comparing the directly calculated psycho-acoustic parameters of every dataset (training, validation and test) with the predicted psycho-acoustic parameters using the trained CNN with the three datasets. The results are presented in Tables V, VI and VII, along with the maximum and minimum values of each parameter in each dataset and the percentual error with respect to the range of the used data.

We can see how, in general, the best results are achieved with the training and validation datasets, as expected, since they have participated in the CNN training process. A good description of the final network performance can be seen in Table VII, which includes an evaluation using the test dataset, which includes data not seen during the network training and validation process. It can be observed that, relatively, the prediction error for the parameters N, S, R and F is greater than the annoyance indicator PA prediction error. These are the best values achieved after testing several CNN architectures and taking into account that the database used is considerably unbalanced as we highlighted in (Figure 1).

TABLE V
PREDICTION RMSE EVALUATION IN TRAINING DATASET

	N	S	R	F	PA
Max. Value	67.27	3.73	25.06	13.37	99.89
Min. Value	0.00	0.24	0.00	0.00	0.00
RMSE Prediction	0.8986	0.1998	0.2783	0.3224	1.5468
% Error	1.33%	5.72%	1.11%	2.41%	1.54%

TABLE VI
PREDICTION RMSE EVALUATION IN VALIDATION DATASET

	N	S	R	F	PA
Max. Value	78.15	3.37	26.97	12.49	97.40
Min. Value	0.00	0.25	0.00	0.00	0.00
RMSE Prediction	0.9292	0.1981	0.4694	0.3622	2.5306
% Error	1.18%	6.36%	1.74%	2.89%	2.59%

TABLE VII
PREDICTION RMSE EVALUATION IN TEST DATASET

	N	S	R	F	PA
Max. Value	46.87	3.08	12.81	6.48	98.62
Min. Value	0.00	0.29	0.00	0.00	0.01
RMSE Prediction	0.8418	0.2320	0.2188	0.3329	1.4013
% Error	1.79%	8.30%	1.70%	5.13%	1.42%

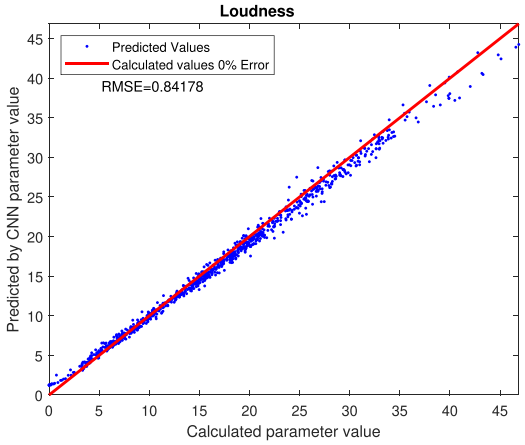


Fig. 4. Calculated N (ground-truth) versus predicted in the the test dataset, specifying as well the final RMSE.

C. Error Distribution

By comparing the results obtained for each parameter, it is observed that as expected, the prediction error tends to be higher for those parameters having an unbalanced range of values in the training dataset, as it is the case of S or F. Nevertheless, the CNN makes very small errors predicting N or R. The prediction PA error is also very low, following our expectations, due to the most influential term for its calculation is N, where the CNN presents very good accuracy.

Some scatter plots of the predicted values versus the calculated ground-truth values are shown in Figures 4, 5, 6, 7 and 8 along with the final RMSE for each parameter, as in Tables V, VI and VII. Comparing now the error distribution within each parameter, it is interesting to discuss the PA behavior (Figure 8) as a summarizing example, since it is assumed to be influenced by the rest of psycho-acoustic parameters. It can be observed that the prediction error tends to increase when predicting high PA values. This is somewhat

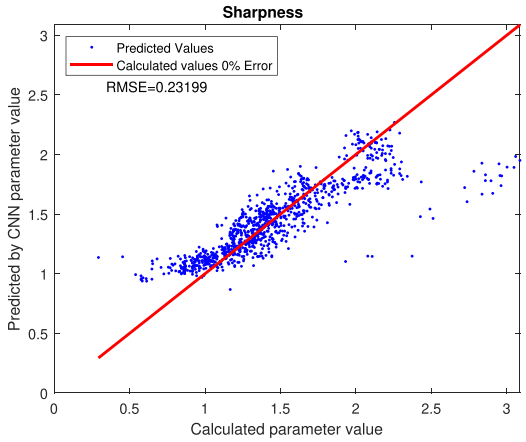


Fig. 5. Calculated *S* (ground-truth) versus predicted in the the test dataset, specifying as well the final RMSE.

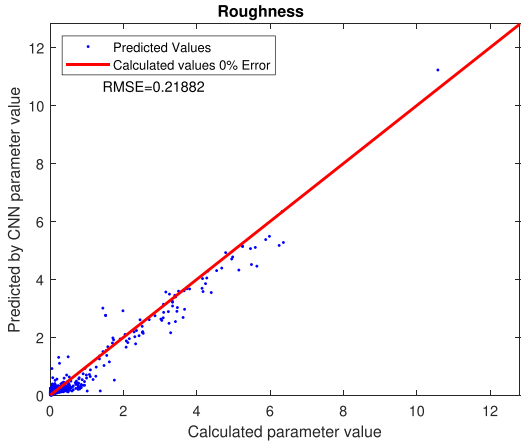


Fig. 6. Calculated *R* (ground-truth) versus predicted in the the test dataset, specifying as well the final RMSE.

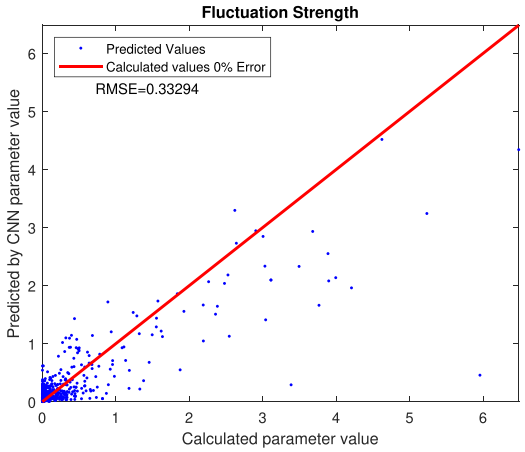


Fig. 7. Calculated *F* (ground-truth) versus predicted in the the test dataset, specifying as well the final RMSE.

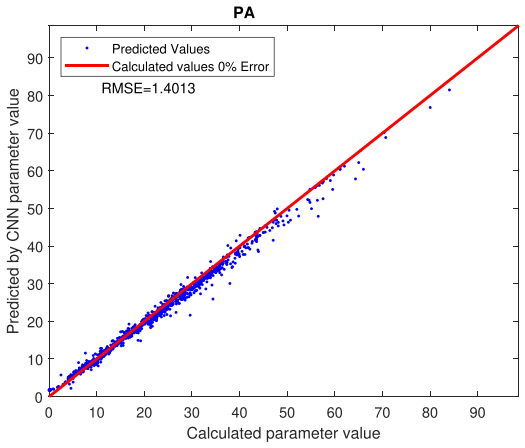


Fig. 8. Calculated *PA* (ground-truth) versus predicted in the the test dataset, specifying as well the final RMSE.

expected, since few examples with PA values above 50 are actually present in the dataset. As a result, the predicted PA tends to be more scattered for values above 50. A similar behavior is observed in general for the other parameters which, taking into account the training data distribution, shows errors that tend to be higher for those audio segments with extreme values. However, notice that, despite these unusual deviations, the CNN model provides parameter predictions and PA predictions with very good accuracy in most cases.

D. Field Test

As most of the audio examples used in the training, validation and testing of the network belong the the UrbanSound8K dataset, an independent test was carried out using new recordings from different urban environments obtained with a Zoom 4HN Pro recorder. This new test is therefore aimed at validating the accuracy of the trained CNN over audio examples (2000 files of 1 second duration) from scenarios not seen during the training phase. The results are shown in Table VIII and it can be observed that they are very similar to those obtained using the validation dataset. This demonstrates the robustness of the CNN, which has been shown to be capable of predicting with good accuracy the psycho-acoustic parameters of the new sound recordings, despite having been acquired in different scenarios and with different audio equipment.

TABLE VIII
PREDICTION RMSE EVALUATION IN INDEPENDENT DATASET

	N	S	R	F	PA
Max. Value	72.32	2.94	6.03	5.58	85.4
Min. Value	0.64	0.00	0.00	0.00	0.00
RMSE Prediction	0.93	0.2403	0.2609	0.2925	2.1704
% Error	1.29%	10.43%	4.32%	5.23%	2.54%

E. Computation Time

This section discusses the reduction in computation time achieved by the proposed CNN with respect to the time required by direct calculation for the studied parameters. Previous works have shown that the number of floating-point operations do not accurately reflect the compute time of CNN-based architectures [29], [30]. Thus, since we are only interested in real-time acoustic monitoring, we center our evaluation in the measured computation time for obtaining the considered psycho-acoustic parameters using direct calculation and neural networks in several architectures. To this end, we measured the time used by the classical algorithms to calculate the 4 psycho-acoustic parameters N, S, R and F

TABLE IX
PSYCHO-ACOUSTIC PARAMETERS COMPUTING VS PREDICTION: TIME COMPARISON

	μ [s]		σ^2 [s^2]	
	PC1	RPi3B+	PC1	RPi3B+
Direct-calculation	1.3052	1.5182	$2.25 \cdot 10^{-3}$	$5.95 \cdot 10^{-3}$
CNN Prediction	0.0050	0.0259	$8.51 \cdot 10^{-6}$	$1.67 \cdot 10^{-4}$

TABLE X
PSYCHO-ACOUSTIC PARAMETERS COMPUTING VS PREDICTION: RAM USAGE COMPARISON

	PC1 (MB)	RPi3B+ (MB)
Direct-calculation	83.81	83.34
CNN Prediction	12.57	18.36

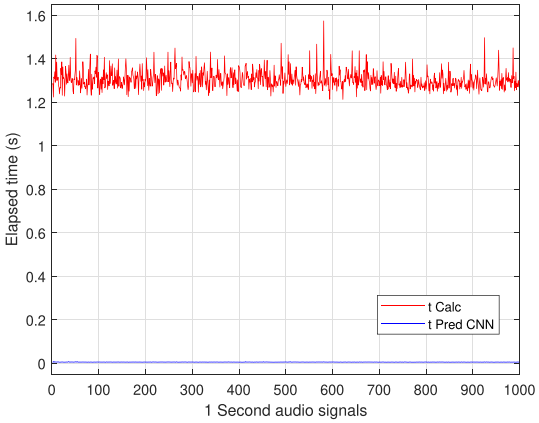


Fig. 9. Elapsed time in direct calculation and elapsed time in prediction using PC1.

and the psycho-acoustic annoyance PA, and the time used by the neural network to predict the five values at once, using the computers and WASNs devices described in Section II-B in both cases. The results are shown in Figure 9, obtained by running the algorithms and CCN on the PC1. Note that we are calculating the parameters and feeding the CNN system with the 1000 audio segments of 1 second duration corresponding to the test dataset.

We see in this figure, that the time required by the CNN to predict all the psycho-acoustic parameters and the PA value (in blue) from raw audio signals is significantly smaller than the one obtained by direct calculation (in red). The average time μ for the CNN prediction is 0.005 s, while for direct-calculation is 1.305 s in the desktop PC1, as specified in Table IX. Thus, a speedup higher than 250 times is achieved by the proposed CNN-based system on this platform, while for the Raspberri Pi3B+, the achieved speedup is close to 60, fulfilling the real-time requirement. The difference between both platforms (PC1 and RPi3B+) is expected due to the limited resources of the Raspberri Pi CPU unsurprisingly. In Figure 9 can be also observed that the times obtained from the CNN are nearly constant, while the times obtained by direct calculation show more dependency on the complexity of the input sound. Table IX specifies as well the measured variance σ^2, reflecting the stability in the computing time required by CNN predictions from raw audio signals.

Besides the required CPU computing time when the parameters are predicted with the CNN or when they are directly calculated, it is also interesting to evaluate RAM usage, especially on wireless acoustic nodes that may have limited resources. The summary of memory usage is shown in Table X in MB, which compares the requirements for direct parameter calculation and for CNN prediction both in a generic PC and in a Raspberry Pi 3B, which is a good example of a WASN device. The values have been extracted from the analysis of the 1000 audio sequences of the test dataset. A significant decrease in RAM usage is observed in both platforms when using CNN prediction instead of direct calculation. It is worth mentioning that such memory savings are especially important in the Raspberry Pi case, since it only has 1GB of memory. Thus, besides the reduction in computation time that makes possible real-time operation, the proposed CNN-based approach allows a more efficient usage of memory (4.5 times less memory than with direct computation).

All these results confirm that the proposed CNN model is a good candidate for IoT-based implementations, where the psycho-acoustic parameters and PA analysis can be comfortably performed in real-time with a computation time much smaller than 1 second.

V. CONCLUSION

In this paper, we proposed a convolutional neural network to estimate the psycho-acoustic parameters and the psycho-acoustic annoyance from raw audio signals more efficiently than using direct calculation. The proposed CNN was shown to be very accurate in predicting the 4 psycho-acoustic parameters (N, S, R, F) and PA value of an input audio signal, quantifying discomfort according to the classical Zwicker's annoyance model. To train such model, we used a large dataset of urban sounds, using the computed annoyance over 1 second segments as ground-truth values for training and validation. Despite having a significantly unbalanced dataset, the resulting model has been confirmed to predict with a very small error in the ranges of the parameters in which more samples were available for training (e.g. PA values in the range going from 0 to 50 PA units), and moderate errors in those audio segments presenting extremely high psycho-acoustic parameters values. On the other hand, we also compared the computing time required by the proposed model and the one obtained by means of direct calculation, obtaining a speedup higher than 250 times, as well as an efficient use of the memory that is critical in these low cost SBC nodes. We have demonstrated the effectiveness of using deep neural networks to estimate the psycho-acoustic parameters in IoT devices with limited resources. This allows to perform the computations in the same node, providing a convenient way to implement smart WASNs aimed at psycho-acoustic assessment more easily and efficiently.

REFERENCES

[1] *Acoustics-Description, Measurement and Assessment of Environmental Noise—Part 1: Basic Quantities and Assessment Procedures*, document ISO 1996-1: 2016, Mar. 2016.
[2] E. N. Directive, "Directive 2002/49/EC of the European parliament and of the council of 25 june 2002 relating to the assessment and management of environmental noise," Tech. Rep., Jul. 2002.

[3] *Acoustics-Soundscape—Part 1: Definition and Conceptual Framework*, document ISO 12913-1, Sep. 2014.
[4] *Acoustics-Soundscape—Part 2: Data Collection and Reporting Requirements*, document ISO/TS 12913-2, Aug. 2018.
[5] M. Cobos, J. Perez-Solano, and L. Berger, "Acoustic-based technologies for ambient assisted living," in *Introduction to Smarte eHealth and eCcare Technologies*, S. Merilampi and A. Sirkka, Eds. Boca Raton, FL, USA: Taylor & Francis Group, 2016, ch. 9, pp. 159–180.
[6] J. Segura Garcia *et al.*, "Spatial statistical analysis of urban noise data from a WASN gathered by an IoT system: Application to a small city," *Appl. Sci.*, vol. 6, no. 12, p. 380, Nov. 2016.
[7] R. Alsina-Pages, F. Alías, J. Socoró, and F. Orga, "Detection of anomalous noise events on low-capacity acoustic nodes for dynamic road traffic noise mapping within an hybrid WASN," *Sensors*, vol. 18, no. 4, p. 1272, Apr. 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/4/1272
[8] F. Alias, R. M. Alsina-Pages, J. C. Socoro, and F. Orga, "DYNAMAP: A low-cost WASN for real-time road traffic noise mapping," in *Proc. FIA-XI Congreso Iberoamericano Acústica; X Congreso Ibérico de Acústica; 49o Congreso Español de Acústica–TECNIACUSTICA 18–24 al 26 de octubre. Cádiz, Sociedad Española Acústica*, 2018. [Online]. Available: http://www.sea-acustica.es/fileadmin/Cadiz18/AAM-5007.pdf
[9] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 6, pp. 2252–2263, Nov. 2006.
[10] M. Hervás, R. Alsina-Pagès, F. Alías, and M. Salvador, "An FPGA-based WASN for remote real-time monitoring of endangered species: A case study on the birdsong recognition of botaurus stellaris," *Sensors*, vol. 17, no. 6, p. 1331, Jun. 2017. [Online]. Available: https://www.mdpi.com/1424-8220/17/6/1331
[11] M. Cobos, J. J. Perez-Solano, S. Felici-Castell, J. Segura, and J. M. Navarro, "Cumulative-Sum-Based localization of sound events in low-cost wireless acoustic sensor networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1792–1802, Dec. 2014.
[12] J. Segura-Garcia, S. Felici-Castell, J. J. Perez-Solano, M. Cobos, and J. M. Navarro, "Low-cost alternatives for urban noise nuisance monitoring using wireless sensor networks," *IEEE Sensors J.*, vol. 15, no. 2, pp. 836–844, Feb. 2015.
[13] H. Fastl and E. Zwicker, *Psychoacoustics: Facts Models*. Berlin, Germany: Springer-Verlag, 2006.
[14] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.
[15] (2018). *Raspberry Pi 3B+*. Accessed: Feb. 1, 2019. [Online]. Available: https://www.raspberrypi.org/
[16] J. Lopez-Ballester, A. Pastor-Aparicio, J. Segura-Garcia, S. Felici-Castell, and M. Cobos, "Computation of psycho-acoustic annoyance using deep neural networks," *Appl. Sci.*, vol. 9, no. 15, p. 3136, Aug. 2019, doi: 10.3390/app9153136.
[17] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Int. Conf. Multimedia - MM*, 2014, pp. 1041–1044, doi: 10.1145/2647868.2655045.
[18] S. A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*. Boca Raton, FL, USA: CRC Press, 2017.
[19] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Amer.*, vol. 71, no. 3, pp. 679–688, Mar. 1982.
[20] H. Lingsong, M. J. Crocker, and Z. Ran, "FFT based complex critical band filter bank and time-varying loudness, fluctuation strength and roughness," in *Proc. Int. Conf. Sound Vib. (ICSV)*, Jul. 2007, pp. 824–832.
[21] A. Pastor-Aparicio *et al.*, "Zwicker's annoyance model implementation in a wasn node," in *Proc. 48th Inter-Noise*, Sep. 2019, pp. 2559–2569.
[22] J. A. Belloch, J. M. Badia, F. D. Igual, and M. Cobos, "Practical considerations for acoustic source localization in the IoT era: Platforms, energy efficiency, and performance," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5068–5079, Jun. 2019.
[23] J. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards End-to-End acoustic localization using deep learning: From audio signals to source position coordinates," *Sensors*, vol. 18, no. 10, p. 3418, Oct. 2018.
[24] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 892–900. [Online]. Available: http://dl.acm.org/citation.cfm?id=3157096.3157196
[25] T. Giannakopoulos, E. Spyrou, and S. J. Perantonis, "Recognition of urban sound events using deep context-aware feature extractors and handcrafted features," in *Artificial Intelligence Applications and Innovations*, J. MacIntyre, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds. Cham, Switzerland: Springer, 2019, pp. 184–195.
[26] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3734–3738.
[27] J. Lee, T. Kim, J. Park, and J. Nam, "Raw waveform-based audio classification using samplelevel CNN architectures," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Dec. 2017, pp. 866–871.
[28] J. Naranjo-Alcazar, S. Perez-Castanos, I. Martin-Morato, P. Zuccarello, and M. Cobos, "On the performance of residual block design alternatives in convolutional neural networks for end-to-end audio classification," 2019, *arXiv:1906.10891*. [Online]. Available: http://arxiv.org/abs/1906.10891
[29] Z. Lu, S. Rallapalli, K. Chan, and T. La Porta, "Modeling the resource requirements of convolutional neural networks on mobile devices," in *Proc. ACM Multimedia Conf. MM*, 2017, pp. 1663–1671, doi: 10.1145/3123266.3123389.
[30] D. Justus, J. Brennan, S. Bonner, and A. S. McGough, "Predicting the computational cost of deep learning models," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 3873–3882.

Jesus Lopez-Ballester received the B.S. and M.Sc. degrees in telecommunications engineering from the University of Valencia, Valencia, Spain, in 2009 and 2014, respectively, where he is currently pursuing the Ph.D. degree in information technology, communications and computation.

He was with the Institute of Robotics, University of Valencia. His current research interests include analysis of acoustic events, e-Health, signal processing, human–machine interfaces, deep learning, and signal processing.

Dr. Lopez-Ballester was a recipient of the Extraordinary Prize Final Project of the College of Telecommunications Engineers.

Adolfo Pastor-Aparicio received the B.S. degree in telematics engineering from the University of Valencia, Valencia, Spain, in 2018, where he is currently pursuing the master's degree in telecommunications engineering with the Department of Computer Science.

He is currently a Researcher with the Department of Computer Science, University of Valencia. His current research interests include the Internet of Things platform and node management, digital audio processing, and virtual network management.

Santiago Felici-Castell received the M.Sc. and Ph.D. degrees in telecommunication engineering from the Polytechnical University of Valencia, Valencia, Spain, in 1993 and 1998, respectively.

He is currently an Associate Professor with the University of Valencia. He is also a Cisco Systems Certificated Instructor, and has authored over 25 technical papers in international journals and conferences. His current research interests include networking, communication systems, and multiresolution techniques for data transmission with quality of service.

Jaume Segura-Garcia received the M.Sc. and Ph.D. degrees in physics from the University of Valencia, Valencia, Spain, in 1998 and 2003, respectively. After completing his Ph.D. study, he was with the Robotics Institute, University of Valencia, where he was involved in several projects related to intelligent transportation systems. Since 2008, he has been with the Department of Computer Science, University of Valencia, where he is currently an Associate Professor. He has been a Visiting Researcher with multiple European research centers. He has coauthored over 90 publications at national and international journals, book chapters, and conferences. He was on the Organizing Committee of several national and international conferences, including the International Workshop on Virtual Acoustics (2011). He is a member of the Spanish Acoustics Society and the European Acoustics Association.

Maximo Cobos (Senior Member, IEEE) received the master's degree in telecommunications and the Ph.D. degree in telecommunications engineering from the Polytechnical University of Valencia, Valencia, Spain, in 2007 and 2009, respectively. He completed his studies with honors under the University Faculty Training Program, and was a recipient of the Ericsson Best Ph.D. Thesis Award on Multimedia Environments from the Spanish National Telecommunications Engineering Association. He received a Campus de Excelencia Postdoctoral Fellowship to work with the Institute of Telecommunications and Multimedia Applications, Valencia, in 2010. In 2009 and 2011, he was a Visiting Researcher with Deutsche Telekom Laboratories, Berlin, Germany, where he worked on the field of signal processing for spatial audio. Since 2017, he has been an Associate Professor with the University of Valencia. His work is focused on the area of digital signal processing for audio and multimedia applications, where he has authored over 70 technical papers in international journals and conferences. He is a Full Member of the Acoustical Society of America.

.

C. Speech Intelligibility Analysis and Approximation to Room Parameters through the Internet of Things

applied sciences

MDPI

Article

Speech Intelligibility Analysis and Approximation to Room Parameters through the Internet of Things

Jesus Lopez-Ballester [1,*,†], Jose M. Alcaraz Calero [2,†], Jaume Segura-Garcia [1,†], Santiago Felici-Castell [1,†], Miguel Garcia-Pineda [1,†] and Maximo Cobos [1,†]

[1] Computer Science Department, Escola Tècnica Superior d'Enginyeria, Universitat de València, 46100 Burjassot, Spain; jaume.segura@uv.es (J.S.-G.); santiago.felici@uv.es (S.F.-C.); miguel.garcia-pineda@uv.es (M.G.-P.); maximo.cobos@uv.es (M.C.)

[2] School of Computing, Engineering and Physical Sciences, University of the West of Scotland, Paisley PA1 1LU, UK; jose.alcaraz-calero@uws.ac.uk

* Correspondence: jesus.lopez-ballester@uv.es

† These authors contributed equally to this work.

Abstract: In recent years, Wireless Acoustic Sensor Networks (WASN) have been widely applied to different acoustic fields in outdoor and indoor environments. Most of these applications are oriented to locate or identify sources and measure specific features of the environment involved. In this paper, we study the application of a WASN for room acoustic measurements. To evaluate the acoustic characteristics, a set of Raspberry Pi 3 (RPi) has been used. One is used to play different acoustic signals and four are used to record at different points in the room simultaneously. The signals are sent wirelessly to a computer connected to a server, where using MATLAB we calculate both the impulse response (IR), and different acoustic parameters, such as the Speech Intelligibility Index (SII). In this way, the evaluation of room acoustic parameters with asynchronous IR measurements two different applications has been explored. Finally, the network features have been evaluated to assess the effectiveness of this system.

Keywords: WASN; room acoustics; impulse response; speech intelligibility index; room parameters estimation

Citation: Lopez-Ballester, J.; Alcaraz-Calero, J.M.; Segura-Garcia, J.; Felici-Castell, S.; Garcia-Pineda, M.; Cobos, M. Speech Intelligibility Analysis and Approximation to Room Parameters through the Internet of Things. *Appl. Sci.* **2021**, *11*, 1430. https://doi.org/10.3390/app11041430

Received: 26 December 2020
Accepted: 3 February 2021
Published: 5 February 2021

1. Introduction

Since Wallace Clement Sabine [1] published his research on reverberation in room acoustics, many researchers have devoted much effort to measuring acoustic behavior in indoor environments. The measurement procedure was standardized in ISO 3382 [2–4], and since then several revisions have been made. This standard makes some recommendations on different aspects related to the measurement procedure (e.g., exciting signal used, number of measurements, etc.), but also leaves some non-mandatory aspects to allow innovation. It also opens the possibility of evaluating other parameters related to the amount of energy from the impulse response in every receiver location.

Wireless Acoustic Sensor Networks (WASN) have allowed some technical automation of the measurement procedure following specific recommendations and following the ISO standard. Further research is needed to improve the measurement procedures by addressing different signal processing techniques [5–9] and the measurement protocols. Using WASN technology we gather the acoustic information of the whole room in one single shot from the sound source using a series of sensors distributed in the room.

In recent years, WASN has been widely applied to different acoustic fields in indoor and outdoor environments. Most of these applications are aimed at locating one or several sources [10,11], tracking [12] or identifying them [13], and measuring specific characteristics of the environment involved [14]. In [15], the authors use a WASN simulation to estimate the room acoustic performance, using localized and averaged Room Impulse Response

(RIRs), and adapt the multipoint equalization of a sound system, to render the sound considering the inversion of the room acoustic model, and applied as a pre-filter to the "dry" signal before playback.

Most parameters used for acoustic analysis of rooms such as Reverberation Time 30 (TR30), TR60 and Clarity 80 and most of the parameters used for speech transmission analysis such as Speech Transmission Index (STI), Clarity 50 and Speech Intelligibility Index (SII) [16] make us think of traditional spaces such as theaters, cinemas, music schools or music production rooms. However, it is interesting and very relevant to make use of such parameters in rooms dedicated to other purposes where the transmission of the word is vital, such as teaching spaces or those dedicated to medical care, to assess their behavior and suitability [17–19] of speech transmission.

The afore mentioned parameters are calculated based on the impulse response of the room with the only exception of the SII [20], which is extracted from a speech signal recorded in the position of the room where we wish to analyze intelligibility. The calculation of parameters based on impulse response is much more complex than SII, so it is not possible to do it in a simple, fast or effective way in a node of our WASN, as it happens for example with the parameters of psychoacoustic disturbance [21]. In fact, we proposed to implement a convolutional neuronal network (CNN) that would allow us to successfully predict them, just as we did with the parameters of psychoacoustic disturbance [21]. Psychoacoustic annoyance quantifies how disturbing different sounds can be to humans. For this purpose, different psychoacoustic parameters (Loudness, Sharpness, Roughness and Fluctuation Strength) are calculated and used to calculate an annoyance marker: Psychoacoustic Annoyance. In this previous work, we used a CNN to perform the prediction of these parameters in a fast and accurate way and to be able to perform the calculations in the node of a WASN itself.

The implementation of SII from a speech signal recorded in any room location we wish to analyze, makes it much easier the necessary signal processing and therefore the calculation capacity of the devices used. Taking into account the current pandemic times and the problems derived from the COVID-19 in the field of education, the adaptation of classrooms originally not intended for teaching has been considered a must to deal with the implementation of social distancing policies (see Figure 1). In this context, we found particularly interesting to be able to have a rapid measurement of speech intelligibility at different points in a room and to be able to act accordingly.

In this work, we present the design and implementation of a WASN system based on RPi, Portaudio library and MATLAB, to perform synchronous and asynchronous acoustic measurements mainly oriented to study the speech intelligibility in a room or SII parameter, as well as the early reflections in the room and the acoustic characteristics of the space.

Although SII can be computed asynchronously, because each node will give a value of SII, the other parameters have been recorded synchronously through the WiFi connection of the RPi. Although this synchronization is not too precise (due to the delays introduced by the network), it is good enough for the purpose of this analysis, which will allow us to have a general idea of the acoustic behavior of the room.

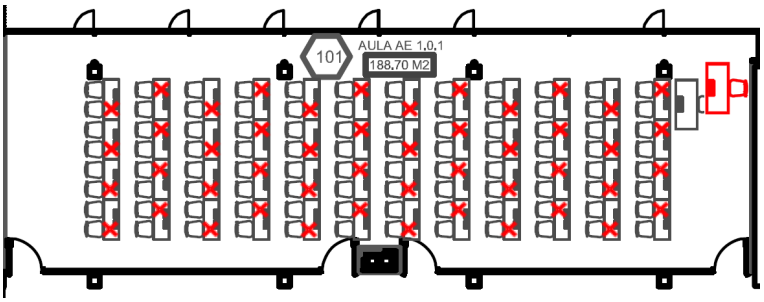


Figure 1. Room schema as an example of social distancing and reduced capacity (red X locations are not occupied).

The paper is structured as follows. Section 1 focuses the problem and gives a slight focus on the related work discussing different approaches and establishing a framework for our proposal. Section 2 addresses the methodology and the materials used to measure the room parameters focused with the planned WASN. Section 3 describes the acoustical measurements performed in a lecture room at the ETSE of the University of Valencia with a discussion and finally, Section 4 gives some conclusion obtained from the deployment process.

2. Materials and Methods

To develop the aforementioned applications, we have used a WASN with 5 nodes based on the Raspberry Pi 3b (RPi) board, 4 for registration and one for emission, according to Figure 2. This platform is based on the Broadcom BCM2837 system on chip, which includes a 1.2 GHz quad-core ARM Cortex-A53 processor, with 512 KB of shared L2 cache, a VideoCore IV graphics processing unit (GPU) at 400 MHz clock frequency, with built-in WiFi (IEEE 802.11 b/g/n standard) and Bluetooth and 1 GB of LPDDR2 RAM at 900 MHz, with a micro-SD card memory. We installed in the measurement nodes an Andoer B01LCIGY8U-USB condenser microphone, with omnidirectional reception pattern, sensitivity = -30 ± 3 dB, SNR > 36 dB and linear frequency response between 20 Hz and 16 KHz.

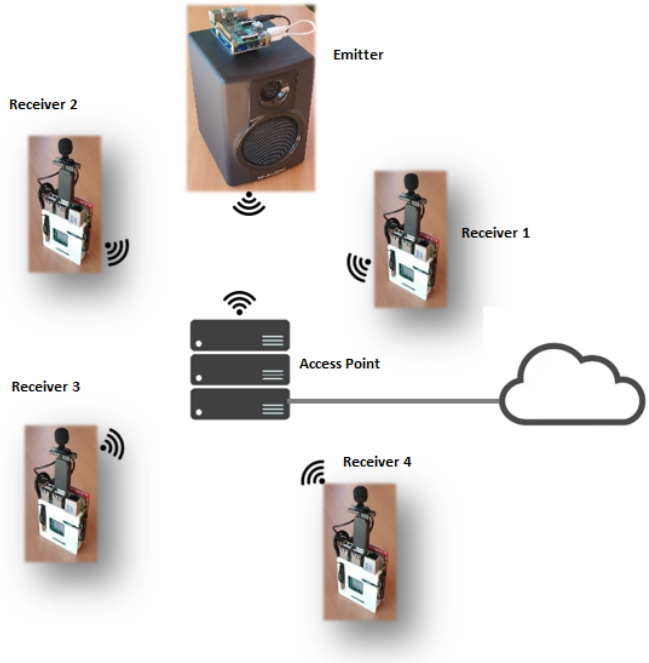


Figure 2. Diagram of the IoT architecture for the room acoustics parameter measurement system.

The set configuration allows each node to be prepared to automatically acquire a configurable audio time with 16 bits per sample (allowing a dynamic range of $20 \cdot log_{10}(2^{16}) = 96.33$ dB) at a sampling frequency of 44.1 kHz. The RPi have been placed in protective structures, each with 3.7 V and 3800 mAh batteries. Measurements have provided us with results of approximately 10 h, more than enough time to perform the analysis, since it is done in minutes. It is very useful to have batteries, since in the case of SII, we can place the nodes in different positions and calculate intelligibility values in more than 4 locations, due to the fact that measurements do not have to be simultaneous. Thus, the sound picked up by the microphone of each node is used to calculate independently the acoustic parameters of the room (focusing on the SII), so it is not necessary in this case to make any data fusion.

Figure 3 shows the classroom prepared for measurement. As shown in Figure 2, the measurement system comprises the registration nodes and the broadcast nodes configured,

with a NTP server configured to have the same time in all the nodes, from a central computer, which acts as a server. The synchronizing order is sent to the player node (with loudspeaker) to start playing the excitation signal and to the recorder nodes to start recording at the same time. In the case of playback, it is possible to play both a frequency sweep that will allow the extraction of the IR of the room and subsequent calculation of different parameters, and different speech signals, which will allow the direct calculation of the SII at the position of each node. Once the recording is finished, the audios are sent to the central equipment which is in charge of storing them along with the positions of each node and making the pertinent computations.

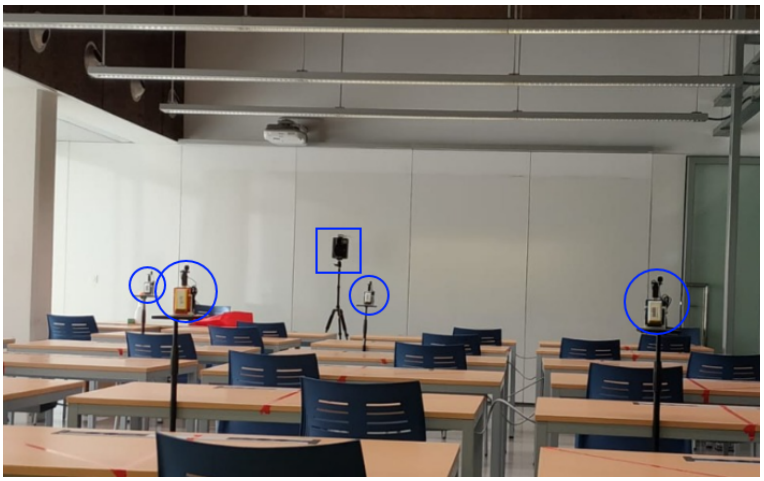


Figure 3. Measurement of parameters in the classroom using the designed system. Detail of sound source location (square mark) and receiver nodes (circular mark).

Audios from 10 s to 1 min duration are used for frequency sweeps, but excellent results have been obtained for SII analysis with 10-s audios and even with 5-s audios of constant speech. The signals used to extract the average impulse response are frequency sweeps from 20 to 20 KHz of 5 and 10 s duration. The human speech signals have been previously recorded in an anechoic environment and belong to 6 professors of the department, 3 male and 3 female in order to have a diverse set of voices.

The software has been designed and tested initially with software simulations of classrooms and rooms of different geometries. Currently, the acquisition and calculation system is operational, and has been tested in different classrooms of the ETSE of the University of Valencia, taking advantage of their adaptation to the interpersonal distance required for teaching.

The parameters evaluated are shown in Table 1. We can calculate only one (SII) from a speech signal alone. Therefore, we will focus the presentation of results on this parameter. The study of SII can give us a clear idea of how the position of the source, distance, geometry and classroom materials affect the intelligibility of the word. Our system as a whole allows measurement of SII during a live class, without the need to interfere in it, only recording the speaker with the nodes located in the desired positions.

SII Computation

The SII value ranges between 0 and 1, and it is highly correlated with speech intelligibility under adverse hearing conditions, such as auditory masking, filters, and reverberation [20]. The American National Standards Institute (ANSI) defined the standard SII to evaluate speech intelligibility in background noise. The measurement inputs are clean speech signal and noise, while the output is a scalar number that specifies the amount of speech intelligibility. A key component of the SII measure is the band importance function, which determines the contribution of each frequency band (e.g., the one-third octave bands) to speech intelligibility. This objective measure is based on the SNRs weighted, usually in

the one-third octave bands. The SII measure is calculated by considering the SNR of each frequency band weighted according to its contribution to speech intelligibility. The SII has a bell-shaped band importance function, as we can see in Figure 4, with a value of 0.0083 for the one-third octave frequency band centered at 160 Hz, 0.0898 for the band centered at 2000 Hz, and 0.0185 for the band centered at 8000 Hz.

Table 1. A set of room acoustic parameter, the source needed for computation and regulation if any.

Parameter	Needed for Computation	Regulation
RT60 (Reverb. Time 60dB)	Impulse Response	ISO 3382-2
C50 (Speech Clarity)	Impulse Response	ISO 3382-2
C80 (Music Clarity)	Impulse Response	ISO 3382-1
STI (Speech Transmission Index)	Impulse Response	IEC 60268-16 2003
SII (Speech Intelligibility Index)	Audio signal	ANSI S3.5-1997

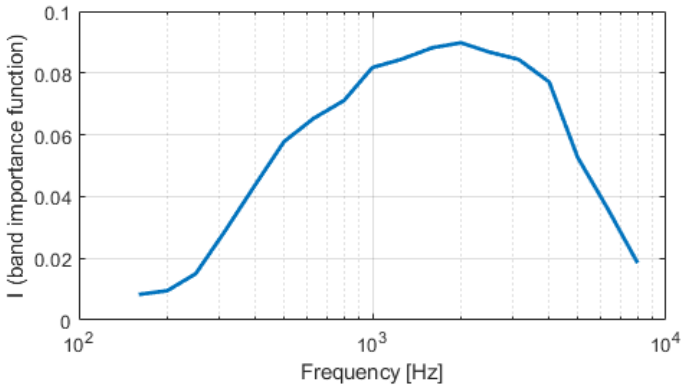


Figure 4. Band Importance function values for average speech SII calculation.

The general formula for calculating the SII is:

$$SII = \sum_{i=0}^{n} I_i A_i \tag{1}$$

where n is the number of individual frequency bands used for computing. Here, there is some flexibility in the definition of the SII standard [20] and it is possible to select the frequency measure bands (ranging from 6 [octave bandwidth] to 18 [critical bandwidth] bands). Generally speaking, the more frequency-specific your measures, the more accurate your computations. The I_i values, also known as the band importance function (BIF), are related to the importance for each frequency band to speech understanding and are based on specific speech stimuli. When they are summed across all bands, they are 1. The standard also allows some flexibility in using the most appropriate BIF for each situation. Finally, the A_i values, or band audibility function (BAF) -also ranging from 0 to 1-, indicate the proportion of speech cues that are audible in a given frequency band and its determination is based simply on the level of the speech, in a given frequency band, relative to the level of noise in that same band. For determining A_i a dynamic range of speech of 30 dB is assumed. Using the basic formula for calculating A_i, it is possible to subtract the spectrum level of noise from the spectrum level of the speech (in dB) in a given band, add 15 dB (the assumed speech peaks), and divide by 30. Resulting values greater than 1 or less than 0 are set to 1 and 0, respectively. This value essentially provides the proportion of the 30-dB dynamic range of speech that is audible to the listener.

Figure 5 provides a complete description of the SII algorithm based on ANSI standard [20], where the BIF are directly defined in the standard and the Equivalent Hearing Threshold levels T_i are the ones used in an average normal audition pattern. This T_i levels will modulate the BAF and combined with BIF, we obtain SII.

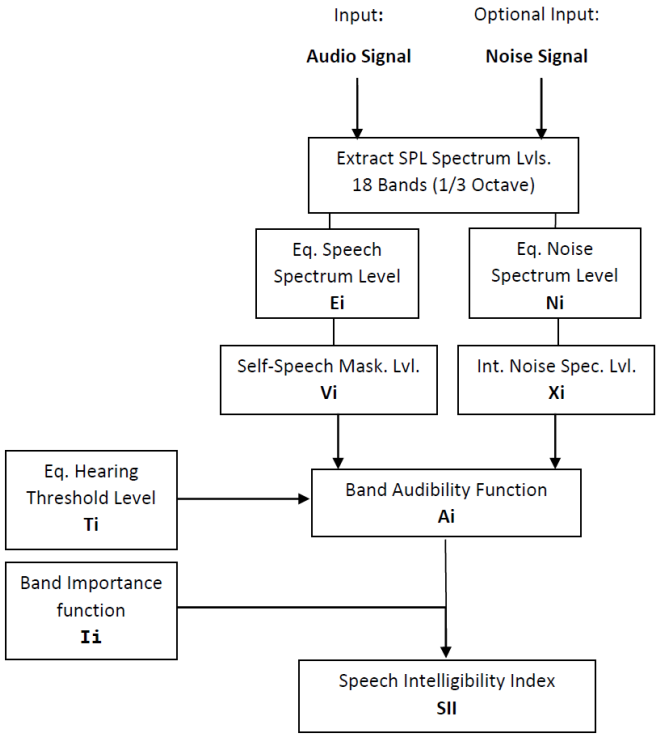


Figure 5. Block diagrams of the SII calculation algorithm defined in ANSI S3.5-1997.

3. Results

The best synchronization that has been achieved with the system described in Figure 2 is 30us, so, although we record at 44.1 kHz our network recording system only allows reliable calculations up to a signal frequency of 16 kHz (as it is the nearest to Nyquist frequency with the synchronized signal), which is enough for the system we are concerned with applied to classrooms and speech, and that has allowed us to focus on SII. In fact, in SII, for example, the interesting thing is to analyze up to 8 kHz in 1/3 octave bands, which is the most precise thing described by the ANSI S3.5-1997 standard [20].

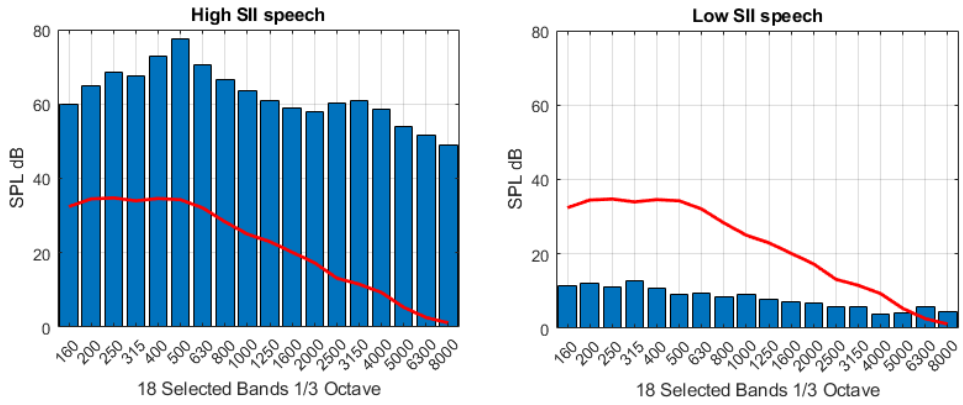
3.1. SII Evaluation

Focusing on SII, we have developed the algorithm following the norm ANSI S3.5-1997. We sampled at 44.1 KHz and from the total audible spectrum of 30 bands in 1/3 octave, we selected the 18 bands specified by the ANSI standard. These bands, as we can see in Figure 6, cover on the frequency content of the human voice and are between 160 and 8000 Hz. The results shown in this Figure describe two examples with high and low SII, taking into account the intelligibility scale from Table 2.

Table 2. Intelligibility scale according to SII parameter.

SII Value	Intelligibility
0.00–0.30	Very bad
0.30–0.45	Bad
0.45–0.60	Acceptable
0.60–0.75	Good
0.75–1.00	Excellent

Figure 6a corresponds to a recording with SII around 0.79 and an excellent perception and Figure 6b shows the frequency bands of an audio corresponding to a room with SII around 0.48 and a bad intelligibility. Figure 6 also shows the Equivalent hearing threshold levels T_i for normal speech used in the SII calculation. The lower the received audio SPL levels are below these levels, the worse will be the speech intelligibility.



(**a**) (**b**)

Figure 6. 18-band spectrum needed for SII computation for: (**a**) excellent intelligibility and (**b**) bad intelligibility. Equivalent hearing threshold levels T_i for normal speech as the overlapping line (red).

As far as we can see in this figure, for the computation of SII, not only the Sound Pressure Levels (SPL) per band are important, but also their distribution in the spectrum, since each band is given a specific weight of importance according to the ANSI standard (the so-called "One-third octave frequency importance functions").

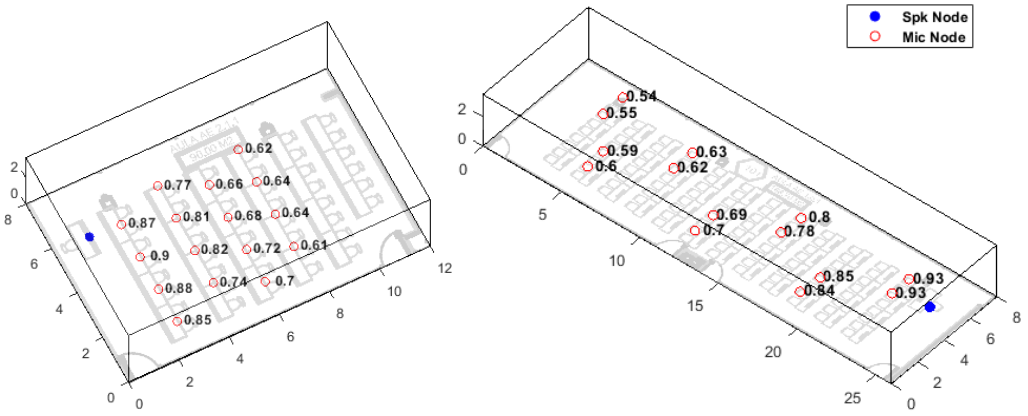
3.2. Study of the Application of the SII Measurement System in a Room

Once the RT60, C50, C80 and STI parameters are determined in the evaluating room, we have applied the intelligibility analysis in different positions in a classroom to plot SII maps according to the position where the hearer is located and therefore, be able to estimate which locations have the worst perception for speech.

A representation system has been developed that reproduces in 2D and 3D the dimensions of the room to be treated, as well as the calculated values of SII for the indicated positions of each node. The system has been prepared to overprint a plan of the room if available to facilitate the analysis of the results.

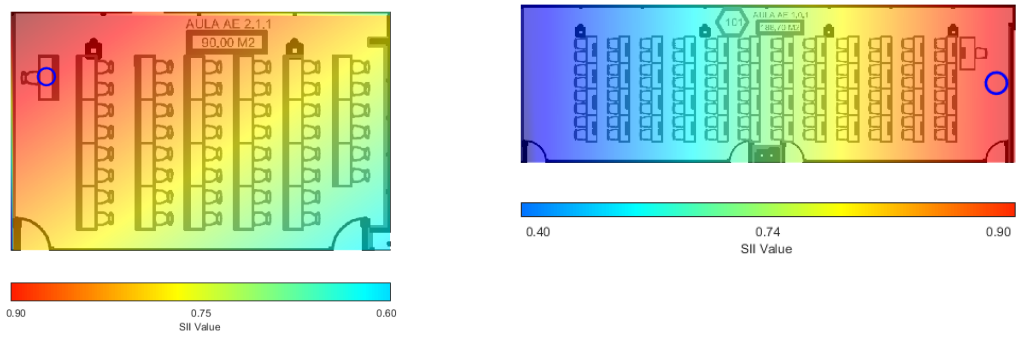
Figure 7 shows two 3D models (out-of-scale) with a SII study in 2.1.1 and 1.0.1 classrooms at the ETSE of the University of Valencia. The dimensions of these rooms are $8 \times 12 \times 3$ and $8 \times 26 \times 3$ meters, respectively. In Figure 8, two heat-map-like (out-of-scale) representations of the SII can be seen, allowing us to graphically assess speaker distances and positions from which there may be intelligibility problems.

Since the SII measurements are asynchronous, we can place the 4 receiver nodes of our system in as many positions as we wish and take numerous samples of the intelligibility by executing the code multiple times. In Figure 7a, we can see the 16 positions where the measurements have been performed with the obtained SII values. On the other hand, in Figure 7b, we can see that 14 measurements have been taken, demonstrating the flexibility of the system when taking SII samples. In Figure 8, we can see 2 SII representations of the same classrooms (2.1.1 and 1.0.1) in the form of a heat-map, which allows a draft of the speech intelligibility in the room, at a glance. The more samples we take, the more accurately we will generate the heat-maps shown in Figure 8.



(**a**) (**b**)

Figure 7. 3D models and SII measurement mappings in different rooms at the ETSE ((**a**). Classroom 2.1.1 and (**b**). Classroom 1.0.1) . Marked positions of sound source node (blue dot) and receiver nodes (black circles).



(**a**) (**b**)

Figure 8. Heat-maps of SII measurement mappings in different rooms at the ETSE ((**a**). Classroom 2.1.1 and (**b**). Classroom 1.0.1). Marked position of sound source node (blue cicle).

The values obtained represent accurately the perception of speech in the rooms studied. In the case of classroom 2.1.1, we have placed the node that simulates the speaker in the position of the teacher's table, as we can see in Figures 7a and 8a, since this is the most common place in this classroom. From the first rows with maximum SII values of 0.9 we see how the SII values decrease until the minimum of 0.61 in the last positions studied. The heat-map in Figure 8a helps us to see more easily how this drop in SII has a diagonal pattern from the speaker position to the opposite corner of the classroom.

The analysis of SII in classroom 1.0.1, shown in Figures 7b and 8b, has been performed by placing the emitter node in a central position with respect to the transverse axis of the classroom, because in this classroom it is common for the speaker to be located in this area.

In Figure 7b we can see that the maximum values of SII are 0.93 while the minimum is 0.54. The values in the areas farther away from the speaker are now lower than in the previous case, a fact that is more easily observed in Figure 8b where darker shades of blue are visible in the background of the classroom. However, as class 1.0.1 is more than twice the size of class 2.1.1, SII would be even lower. We have found the explanation in the reverberation that reinforces the acoustic signals from the emitter and allows the SPL levels to not drop as much towards the end of the classroom as would be expected. Specifically, RT60 measured in classroom 2.1.1 is 0.58 s while RT60 in classroom 1.0.1 is

0.86 s, which explains the reinforcement of the acoustic signals towards the back of the room, due to reflections.

This representation is launched automatically as soon as the audios are received and the parameters are calculated, so you can put the system to record and automatically show the analysis and store it in a matter of seconds, which represents a saving in time and simplicity when evaluating rooms.

4. Conclusions

In this work has been designed, prototyped and deployed a IoT system for SII measurement, from the ANSI standard for acoustic assessment of rooms. At present, and due to the necessary social distance that has been implemented in schools, it is of particular interest to have a system such as the one designed to assess speech intelligibility in different spaces. It is true that there are systems on the market that offer similar analyses, at a very high price compared to the presented system, and with a technological complication also excessive for a first and quick analysis.

The complexity of the system has been reduced to a minimum so that the user does not have to worry about the technological part and only focuses on the positions he wants to analyze in each room. Likewise, to facilitate this task the nodes of our system are totally autonomous in terms of energy and communicate wirelessly, so it facilitates its deployment and location in any area. The designed system implements a data visualization in a clear, clean and fast way, since in a matter of seconds it allows graphical depiction of the measured results of SII allowing their evaluation by the user.

In addition, the system can store the desired data in a cloud server and be able to retrieve and display it at any time. Thanks to this, different speech intelligibility improvement techniques can be tested, such as those based on loudspeakers hardware systems and those based on signal processing software, and the results obtained can be measured and compared easily and quickly. All this allows analysis and improvement of speech intelligibility not only in teaching or lecture rooms, but also in any type of enclosure where speech transmission is critical, such as operating theaters, airport control rooms or hospital waiting rooms.

Author Contributions: Conceptualization, J.L.-B. and J.S.-G.; methodology, M.C.; software, J.L.-B.; validation, J.M.A.C.; formal analysis, M.C.; resources, S.F.-C. and M.G.-P.; writing—original draft preparation, J.L.-B. and S.F.-C.; writing—review and editing, J.M.A.C. and M.G.-P.; funding acquisition, J.S.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been partially funded by the Ministry of Innovation and Economy within the project with reference BIA2016-76957-C3-1-R (co-financed with FEDER funds) and the grant with reference BES-2017-082340, and by the Generalitat Valenciana, with grants BEST/2020/117 and AEST/2020/048.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Sabine, W.C. *Collected Papers on Acoustics*; Peninsula Publishing: Los Altos, CA, USA, 1992
2. ISO 3382-1:2009. *Acoustics–Measurement of Room Acoustic Parameters–Part 1: Performance Spaces*; International Standard Organization: Geneva, Switzerland, 2009.
3. ISO 3382-2:2008. *Acoustics–Measurement of Room Acoustic Parameters – Part 2: Reverberation Time In Ordinary Rooms*; International Standard Organization: Geneva, Switzerland, 2008.
4. ISO 3382-3:2012. *Acoustics–Measurement of Room Acoustic Parameters–Part 3: Open Plan Offices*; International Standard Organization: Geneva, Switzerland, 2012.

5. Schroeder, M.R. Integrated-Impulse Method Measuring Sound Decay without Using Impulses. *J. Acoust. Soc. Am.* **1979**, *66*, 497–500. [CrossRef]
6. Farina, A. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In Proceedings of the AES 108th Convention, Audio Engineering Society, Paris, France, 19–22 February 2000; Preprint 5093.
7. Vorlander, M.; Kob, M. Practical aspects of MLS measurements in building acoustics. *Appl. Acoust.* **1997**, *52*, 239–258. [CrossRef]
8. Stan, G.-B.; Embrechts, J.-J.; Archambeau, D. Comparison of different impulse response measurement techniques. *J. Audio Eng. Soc.* **2002**, *50*, 249–262.
9. Mommertz, E.; Muller, S. Measuring Impulse Responses with Digitally Pre-emphasized Pseudorandom Noise Derived from Maximum-Length Sequences. *Appl. Acoust.* **1995**, *44*, 195–214. [CrossRef]
10. Cobos, M.; Perez, J.J.; Felici, S.; Segura, J.; Navarro, J.M. Cumulative-sum-based localization of sound events in low-cost wireless acoustic sensor networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *12*, 1792–1802. [CrossRef]
11. Alexandridis A.; Mouchtaris A. Multiple sound location estimation and counting in a Wireless Acoustic Sensor Network. In Proceedings of the 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 18–21 October 2015.
12. Malhotra, B.; Nikolaidis, I.; Harms, J. Distributed classification of acoustic targets in wireless audio-sensor networks. *Comput. Netw.* **2008**. [CrossRef]
13. Duarte M.F.; Hen, H.Y. Vehicle classification in distributed sensor networks. *J. Parallel Distrib. Comput.* **2004**, *64*, 826838. [CrossRef]
14. Pastor-Aparicio, A.; Segura-Garcia, J.; Lopez-Ballester, J.; Felici-Castell, S.; García-Pineda, M.; Pérez-Solano, J.J. Psychoacoustic Annoyance Implementation with Wireless Acoustic Sensor Networks for Monitoring in Smart Cities. *IEEE Int. Things J.* **2020**, *7*, 128–136. [CrossRef]
15. van Waterschoot, T.; Moonen, M. Distributed estimation and equalization of room acoustics in a Wireless Acoustic Sensor Network. In Proceedings of the 20th European Signal Processing Conference (EUSIPCO 2012), Bucharest, Romania, 27–31 August 2012.
16. Larm, P.; Hongisto, V. Experimental comparison between speech transmission index, rapid speech transmission index, and speech intelligibility index. *J. Acoust. Soc. Am.* **2006**, *119*, 1106–1117. [CrossRef] [PubMed]
17. Lam, C.L.C. *Improving the Speech Intelligibility in Classrooms*; Department of Mechanical Engineering, The Hong Kong Polytechnic University: Hong Kong, China, 2010.
18. McNeer, R.R. Bennett, C.L.; Horn, D.B.; Dudaryk, R. Factors affecting acoustics and speech intelligibility in the operating room: Size matters. *Anesth Analg.* **2017**, *124*, 1978–1985. [CrossRef] [PubMed]
19. Ryherd, E.E.; Moeller, M., Jr.; Hsu, T. Speech intelligibility in hospitals. *J. Acoust. Soc. Am.* **2013**, *134*, 586–595. [CrossRef] [PubMed]
20. American National Standards Institute; Acoustical Society of America. *Standards Secretariat. American National Standard: Methods for Calculation of the Speech Intelligibility Index*; American National Standards Institute: New York, NY, USA, 1998.
21. Lopez-Ballester, J.; Pastor-Aparicio, A.; Felici-Castell, S.; Segura-Garcia, J.; Cobos, M. Enabling Real-Time Computation of Psycho-Acoustic Parameters in Acoustic Sensors Using Convolutional Neural Networks. *IEEE Sens. J.* **2020**, *20*, 11429–11438. [CrossRef]

.

D. AI-IoT Platform for Blind Estimation of Room Acoustic Parameters Based on Deep Neural Networks

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3203570

AI-IoT Platform for Blind Estimation of Room Acoustic Parameters Based on Deep Neural Networks

Jesus Lopez-Ballester, *Graduate Student Member, IEEE,* Santiago Felici-Castell, Jaume Segura-Garcia, and Maximo Cobos, *Senior Member, IEEE*

***Abstract*—Room acoustical parameters have been widely used to describe sound perception in indoor environments, such as concert halls, conference rooms, etc. Many of them have been standardized and often have a high computational demand. With the increasing presence of deep learning approaches in automatic monitoring systems, wireless acoustic sensor networks (WASNs) offer great potential to facilitate the estimation of such parameters. In this scenario, Convolutional Neural Networks (CNNs) offer significant reductions in the computational requirements for in-node parameter predictions, enabling the so-called Artificial Intelligence-Internet of Things (AI-IoT). In this paper, we describe the design and analysis of a CNN trained to predict simultaneously a set of common room acoustical parameters directly from speech signals, without the need for specific impulse response measurements. The results show that the proposed CNN-based prediction of room acoustical parameters and speech intelligibility achieves a relative error rate of less than a 5.5%, accompanied by a computational speedup factor close to 250 with respect to the conventional signal processing approach.**

***Index Terms*—Internet of Things, Room acoustic parameters, speech intelligibility, convolutional neural networks, wireless acoustic sensor networks**

I. Introduction

The prediction of acoustical impression is an issue of major interest in architectural design [1]. The interest in obtaining an accurate acoustic description or monitoring of the sound impression is especially relevant when the analyzed space, whether indoor or outdoor, is intended for applications where the perception of musical or speech sources plays a central role. Underlying such interest, there is a need to understand and control better how sound is perceived by a human listener in a room, but subjective measurements are laborious and expensive. In practice, objective parameters, commonly called room acoustic parameters, are used (e.g. reverberation time, clarity or definition, to name a few). Additionally, other descriptors related to speech intelligibility can be also considered to complete the acoustical description of sound perception in the analyzed space. These parameters are standardized by ISO 3382 and its different revisions [2], [3], [4] for acoustic parameters related to reverberation and energy and also ISO 9921 and ANSI S3.5-1997 [5], [6] for those related to speech intelligibility.

The authors are with the Department of Computer Science, ETSE, Universitat de València, 46100 Burjassot (Valencia), Spain (e-mail: jesus.lopez-ballester@uv.es; santiago.felici@uv.es; jaume.segura@uv.es; maximo.cobos@uv.es).

In the architectural design process, techniques based on virtual models are used to obtain a reasonably controlled acoustic behavior [7][8]. However, such behavior should always be verified by in situ measurements, since any modification of materials, geometry, furniture or even the seating capacity can produce significant changes. It is therefore necessary to have a system that allows the acoustic parameters of the room to be evaluated quickly and easily. Among the wide variety of acoustic parameters included in the standards, a discrete set of parameters is usually selected to fit the analysis that interests most, allowing a complete description of the room under study. There are different commercial products that perform the study and analysis of room acoustical or speech intelligibility parameters, but they are usually expensive and sometimes limited to a very specific use [9], [10].

In this context, the use of WASN can add significant value, since they can collect information and perform measurements in a distributed manner and can also be adapted or reconfigured to acquire further measurements, as well as to incorporate other new parameters. This network is composed of several nodes that have their own processing unit, memory and wireless communication modules (usually WiFi), able to connect microphones and loudspeakers. In general, these networks have usually a low cost and facilitate the expansion of the number of nodes according to the needs or tests to be performed. Nodes are typically based on generic Single Board Computers (SBC), such as the popular Raspberry Pi (RPi) [11]. The nodes that connect to loudspeakers can reproduce different types of sounds (specific tones, sweeping frequencies, background noise, etc.) depending on the parameter under consideration. In the same way, the nodes connected to microphones can record audios with different sampling frequencies and bit depth per sample. These networks are used for different tasks, for instance to locate sound sources [12], track them [13], identify them [14] or measure specific characteristics of the surrounding environment [15], to name some of them.

However, due to the moderate performance of these nodes and the intense signal processing required for the calculation of most acoustic parameters (in addition to the necessary audio sampling and storage), it is necessary to devise alternative implementations for their calculation. In this scenario, neural networks, and in particular deep Convolutional Neural Networks (CNNs), offer significant reductions in the computational requirements for parameter prediction. While deep neural networks can be computationally expensive to

train, once the process is finished and the final weights are available, inference can be really fast, as the computational requirements for estimating the acoustic parameters are only those corresponding to the execution of one iteration with the network, which in our case allows obtaining the parameters in a single forward pass.

In this paper, we propose the combined use of deep learning and IoT technology to gather audio information and to enable the dynamic and fast prediction of acoustical parameters at a reduced computational cost, with a twofold contribution. On the one hand, while other approaches for blind estimation of room acoustic parameters exist, the proposed approach provides a joint prediction of parameters with an end-to-end model accepting only raw speech as input. On the other hand, the model allows for in-situ predictions with considerable computational savings that facilitate acoustic monitoring deployments, providing as well interesting features in terms of power consumption savings and enhancements in security and privacy issues.

The model proposed in this paper was trained and validated using a dataset of sound recordings obtained from synthetic and real room impulse responses (RIRs). Synthetic RIRs were generated considering a variety of rooms with different sizes, microphone configurations and wall-reflection factors, allowing the trained networks to generalize to a wide range of acoustic environments. The resulting CNN model allows predicting the whole set of parameters by exploiting the ability of CNNs to obtain useful internal representations that can be shared for the calculation of different acoustic descriptors. This artificial intelligence model running over a connected IoT system enables what is called AI-IoT. Therefore, this article provides a relevant example of an AI-IoT system for acoustic monitoring, describing a full methodology that goes from data generation and model design to a final in-situ deployment.

The rest of the paper is structured as follows. Section II shows the related work. In Section III, we analyze and describe the traditional room acoustic and speech intelligibility parameters. In Section IV, we present the background and design considerations related to the use of CNNs to estimate blindly these parameters from speech signals and reducing the cost with respect to the direct signal processing computation approach. In Section V, we describe the characteristics and operation of the IoT system we have designed to monitor room acoustic parameters by performing in-node predictions. In Section VI, we show the experimental evaluation and discuss the results. Finally in Section VII, we conclude the paper.

II. Related work

Other recent works have also explored the methodology of reducing the computational cost of complex algorithms to be implemented in IoT nodes within WASN applications. In [16], the authors describe an IoT system for sound recording and direct calculation of psychoacoustic annoyance parameters at the Edge. Also in [17], the same authors present CNN results applied to the calculation of the subjective annoyance, based on the estimation of a set of psycho-acoustic parameters, all of them estimated using a deep learning model with an error probability lower than 3%, allowing to perform calculations locally in the nodes.

In addition, we can find different CNN designs applied to estimate different room acoustic parameters. In [18], a CNN is used to estimate blindly the volume of rooms from reverberant single-channel speech signals in the presence of noise. In a similar way, in [19], a CNN is proposed to estimate the geometry of a room and reflection coefficients given RIRs, achieving an accuracy of 6.5 cm for each dimension in room geometry estimation and 0.09 accuracy in reflection coefficients. In this case, they do not require the knowledge of the relative positions of sources and receivers in the room. Also, in [20], it is shown how deep learning can use the effect of reverberation on speech, to classify a recording in terms of the room in which it was recorded, with a maximum accuracy of 90%.

Works presented in [21], [22] describe CNN models capable of predicting reverberation time ($RT60$) and speech clarity ($C50$) respectively, but on an individual basis. In [23], it is presented a method for blindly estimating the Speech Transmission Index (STI), again individually, without measuring or modeling the impulse response of the room, using deep CNNs and simulated RIRs combined with clean speech examples. The average error in STI prediction was shown to be below 4%. Although there are not many studies aimed at simultaneously predicting several acoustic parameters with the same CNN model, we can find 2 works with this approach. On one hand [24], in order to estimate the reverberation time and the direct-to-reverberant ratio using CNNs, the authors propose a method to expand a small dataset of real RIRs. With that, the performance of the model is improved in terms of speed, being 4-5 times faster. On the other hand, in [25] a wider set of parameters is considered, including $RT60$, early decay time (EDT), musical clarity ($C80$), definition ($D50$) and STI, and demonstrating the capabilities of CNNs to predict a set room acoustic parameters.

Room acoustic parameters are also interesting with regards to analyzing indoor spaces where speech transmission is vital, such as teaching spaces or those dedicated to other issues, for instance medical operating rooms. The acoustic parameters can be analyzed in such spaces to properly assess their acoustic behavior and adequacy [26], [27], [28]. Also, other works must be mentioned that focus on different signal processing techniques for the estimation and measurement of room acoustic parameters [29], [30], [31], [32].

Finally, with regard to the acoustic models, due to the large number of room acoustic parameters, in practice they are combined to highlight sound acoustic behaviors, focused on specific nuances. In particular the most well known models are those described in [7] and [8]. These studies, focused on concert halls, define acoustic models that include several parameters which when taking the appropriate values, together provide an acoustic behavior suitable for the intended purpose, such as the perception of music or voice.

The previous works confirm that the estimation of room acoustical parameters and the prediction of speech intelligibility are classic problems that have both attracted the interest of the research community for a long time, and which continue

to demand innovative and challenging solutions for the AI-IoT era, as the one proposed in this paper.

III. Room acoustic and speech intelligibility parameters

Most room acoustic parameters are standardized by ISO 3382 and its different revisions [2], [3], [4]. The standard also makes some recommendations and specifies procedures on different aspects related to the measurement process, but also leaves some non-binding aspects to allow for innovation.

Thus, different researchers select some acoustic parameters or others depending on the orientation of the study, such as the Gottingen school [33], [34] that considers the interaural cross-correlation index ($IACC$), reverberation time (RT) and clarity (C) or Yamamoto and Suzuki [35] that replaces RT by strength (G). Based on the aforementioned studies and those performed by S. Cerdá and A. Gimenez [36], we have selected a set of 5 parameters of 3 different independent meaningful natures, energetic distribution, reverberation and speech intelligibility. Initially, we selected reverberation time ($RT60$), musical clarity ($C80$) and speech transmission index (STI) as the main independent parameters. Later, practical reasons derived from COVID-19 compelled us to carry out acoustic measurements in rooms enabled for teaching activities when they were not designed for that purpose, so we included speech clarity ($C50$) and speech intelligibility index (SII) to orient our set of parameters to the evaluation of the transmission of oral information.

Table I summarizes the complete set of parameters used in this work along with some details on their processing requirements.

According to subjective impression aspects, these parameters can be grouped into:

- Reverberation time parameters, that represent the degree of vivacity of the hall: $RT60$ (60dB) measures the time it takes for the sound energy to decay by 60 dB. In practice, these measurements typically span the frequency range from 50 Hz to 8 kHz in 1:1 or 1:3 octave bands.
- Energy parameters, that describe the relationship between the energy in early sound reflections with that of late sound reflections. Speech Clarity ($C50$) and Music Clarity ($C80$) are based on the energy in dB of the sound pressure level received within the first 50 ms for $C50$ or 80 ms for $C80$, minus the remaining energy of the signal. The final value for each parameter is based weighting these parameters at different frequencies: 0.5, 1, 2, 4 kHz.
- Intelligibility parameters, that are a measure of how comprehensible speech is in certain conditions as follows:
  - STI (Speech Transmission Index): STI is calculated as the weighted sum of the Modulation Transfer Index (MTI), one for each octave frequency band from 125 Hz to 8 kHz, where each MTI value is obtained from the Modulation Transfer Function (MTF) [37] values over 14 different modulation frequencies, taking into account auditory effects according to IEC 60268-16 to include the effects produced by the reverberation and the noise over the impulse response. STI values varies from 0 = bad to 1 = excellent. On this scale, an STI of at least 0.5 is desirable for most applications [38], [39].
  - SII (Speech Intelligibility Index) is based on the sum of weighted (per band frequency and distortion factor) spectrum level of equivalent speech signal levels and the equivalent disturbance spectrum levels in the octave band, with six middle frequencies from 250 Hz to 8 kHz for a fast octave approach. In our case, for its calculation we have used the most precise scale contemplated by the standard, in 1/3 octave, between 160 Hz and 8 kHz (18 bands), where the SII is calculated as the average of the product between the Importance frequency band (I_i) and the band Audibility (A_i) for each band (i). As in the case of STI, we will obtain a value in the range from 0 (very bad) to 1 (very good intelligibility).

As it is shown in Table I, all the mentioned parameters are calculated from the impulse response of the room, except the SII [40], which can be extracted from a speech signal recorded at the position of the room where speech intelligibility is to be analyzed. The study of SII provides us with a better idea of how the position of the source, distance, geometry and room materials affect speech intelligibility. The proposed AI-based framework will directly estimate all the above parameters from captured speech signals within a single CNN architecture. Although more parameters could have been included, such as the Early Decay Time (EDT) or Brilliance (Br) within the reverberation group, or the or Sound Strength Factor (G) within the energetic ones, we have selected only these 5 parameters for being widely used and sufficiently representative. Thus, this set of parameters will be considered to train and evaluate the performance of the system, which is intended to predict with a low computational cost the whole set of parameters at once, from actual speech recordings within a room, avoiding the need for cumbersome RIR measurements involving test signals such as frequency sweeps or Maximum Length Sequences (MLS).

IV. Convolutional Neural Network for Parameter Prediction

CNNs have proven to be very useful in predicting acoustic-related parameters in a fast and accurate way [17]. This is due to the ability of CNNs to learn optimized signal representations during the training process. We can schematically define a neural network as a set of layers formed by neurons. With a proper training, the combination of layers may learn a mapping from inputs to outputs provided in a training set. CNNs are characterized by convolutional layers that implement different filtering processes. The specific filters applied over the input at each layer are defined during the training process by adjusting the network weights, minimizing the value of some defined loss function. This process allows CNNs to provide internal signal representations derived from a set of network weights representing learned filters optimized to solve a specific task. In an end-to-end network such as the one proposed in this work, the input to the model is a raw audio segment, and

TABLE I
ROOM ACOUSTIC AND SPEECH INTELLIGIBILITY PARAMETERS, DESCRIPTION, CLASSIFICATION AND REQUIREMENTS

Parameter	*Classification*	*Requirements*	*Regulation*
$RT60$ (Reverb. Time 60dB)	Reverberation time	Impulse Response	ISO 3382-2
$C50$ (Speech Clarity)	Energy Speech Clarity	Impulse Response	ISO 3382-2
$C80$ (Music Clarity)	Energy Music Clarity	Impulse Response	ISO 3382-1
STI (Speech Transmission Index)	Intelligibility	Impulse Response	IEC 60268-16, ISO 9921
SII (Speech Intelligibility Index)	Intelligibility	Audio Signal	ANSI S3.5-1997

it is the network that must analyze and extract meaningful features throughout its successive stages. This process leads to meaningful signal representations resulting from a diversity of learned filters, which may represent low and high frequency decompositions and increasing and decreasing amplitudes [41]. Note that other architectures like feed-forward fully-connected networks are not suitable for working with raw audio inputs, as their use makes more sense when being applied to a set of pre-extracted audio features (usually, general-purpose hand-crafted features). In this context, the filters making up convolutional layers have a length corresponding to the local receptive field of a single unit within the layer. The action of such filters upon the input signal creates the output of the layer or feature map. Important aspects of CNNs are those related to parameter sharing (sharing of weights by all neurons in a particular feature map) and local connectivity (each neuron is connected only to a subset of the input nodes). Besides reducing the number of model parameters, this helps to maintain the same feature detector across different sections of the input data.

Despite the fact that CNNs are mostly known for their capabilities to perform signal classification tasks, especially in the image domain, the above parameters vary along a continuous scale. Therefore, we address the CNN design for parameter prediction as a regression problem. Thus, the objective is to obtain an accurate end-to-end architecture capable of predicting the whole set of parameters from raw audio inputs.

A. Dataset

The training of the network requires a large amount of audio data accompanied with annotated ground-truth labels, which in our case corresponds to the actual values of the 5 considered parameters. Moreover, since a high generalization power is desired, the training data must contain recordings corresponding to rooms of very different characteristics. Note that finding a high number of rooms with the appropriate geometric and constructive characteristics to generate enough signal diversity is a complicated task. Thus, a discrete set of impulse responses corresponding to 15 different rectangular rooms with $RT60$ values ranging from 0.1 to 1.5 seconds, has been synthetically generated by means of the image-source method [42], [43]. The size of the room, the reflection coefficients and the source to sensor distance have been adjusted to get different representative $RT60$ values. Moreover, for each synthetic room, a set of 20 RIRs were generated according to different source-microphone configurations considering a rectangular grid. To complete the set of impulse responses, we included responses from 10 more real rooms extracted from the OpenAir repository (Open Acoustic Impulse Response Library) [44], which is an impulse response repository intended for auralization purposes. The responses were selected so that they were measured in indoor spaces. Thus, a total of 310 RIRs from 25 rooms have been used to create a dataset of speech signals recorded in different acoustic environments. While the number of different rooms used in the dataset generation process is moderate, it was verified that this was sufficient to provide state-of-the-art accuracy, as discussed in Section VI.

The speech samples used to generate the dataset were extracted from the DARPA TIMIT acoustic-phonetic continuous speech database [45], which contains 6300 recordings of people pronouncing different phrases. The pre-processing steps applied to the extracted speech signals are as follows: a) the sample speech signals are resampled to 16 kHz to have uniform sampling rates accross the dataset, b) convolve with the different RIRs, c) randomly cut the result to a final duration of 3.5 seconds (56000 samples). To perform the training and assessment of the neural network, the full dataset was divided into three 3 dataset partitions: training, validation and test. The partitions that participate in the training process are only the training and validation ones. For these two datasets, we have used 30000 signals, 80% for training (26000 signals) and 20% for validation (4000 signals). Additionally, we have added a set of 1000 signals which form the test partition, used to test the CNN performance more thoroughly. This last test partition is composed of speech signals that did not participate in the training/validation process convolved with 15 impulse responses that did not participate either (10 real extracted from OpenAir and 5 synthetic) and belonging to different rooms. Similar considerations were taken with respect to the selection of the dry speech signals used to generate the examples in the partitions, ensuring that the speech files used in validation and test correspond to speakers different from the ones in the training dataset.

Preliminary tests during the design process showed that the final accuracy obtained depended much more on the variability in the content of different audios belonging to the same room condition than on the variability of impulse responses with similar room parameters. Nevertheless, to prevent overfitting of the model to data generated from synthetic impulse responses, the training data was balanced by including a similar number of real impulse responses.

B. Design, Configuration and Training

The design of the CNN is based on our previous work [17], where we trained a CNN to predict parameters related to psychoacoustic annoyance. The proposed network is similar

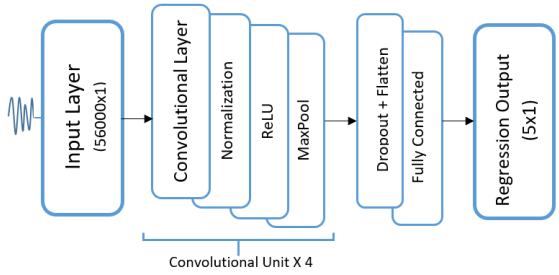


Fig. 1. Proposed CNN architecture and detail of its layers.

to other well-known CNN-based approaches like AlexNet [46] in the image field. This CNN is based on a scheme comprising several convolutional layers followed by a max pool layer, where the size of the filters will decrease, while their number will increase as the network goes deeper. The final configuration in terms of number of layers, kernel sizes and number of filters, in addition to the training parameters defined for our CNN model, corresponds to the best performing model achieved after numerous analyses and tests in the validation partition always taking into account the complexity constraints.

Figure 1 describes the CNN architecture used in this work. The model is based on 4 *Convolutional* stages (from S1, stage 1, to S4, stage 4), each one consisting of a temporal convolution layer, followed by a *Rectified Linear Unit (ReLU)* activation layer that eliminates negative values by setting them to zero, and a *Max-pool* layer that keeps the maximum from a set of elements, implementing a downsampling of the layer input. As mentioned, this convolutional unit (formed by 4 layers) is repeated 4 times. A *Dropout* layer with a dropout probability of 0.3 is used to prevent overfitting, followed by a *Flatten* layer that converts the multidimensional output to a one-dimensional feature vector. Finally, a *Fully connected* layer is followed by an *Output Regression* layer intended to minimize the MSE (without weighting) between the true and predicted values for the 5 considered parameters: $RT60$, $C50$, $C80$, STI and SII. The complete description of these inner CNN configuration parameters that define each layer are shown in Table II.

The training was performed using Adam optimizer with an initial learning rate of 10^{-3}. The maximum number of training epochs was set to 350, with input shuffling considering minibatches of 512 audio signals. The loss function to be minimized during the training process is the mean squared error (MSE), which is obtained from the mean squared differences between the true and predicted values. MSE is very sensitive to outlier values that differ from the mean. Therefore, the MSE is well fitted for regression-based problems where the distribution of the output conditioned on the input data is expected to be Gaussian, and we intend to penalize more (quadratically) larger error values than small ones. This allows us to minimize larger errors with more priority over smaller ones during the training process. As a result, the CNN is adapted to the end-use where large errors in the prediction of the acoustic parameters would lead to an erroneous analysis of the behavior of the analyzed room.

TABLE II
CNN LAYERS DEFINITION

Layer	Size	Filters	Stride
Input	56000×1		
Convolutional S1	512×1	10	10
ReLU. S1			
Max Pool S1	2×1		2
Convolutional S2	256×1	20	5
ReLU. S2			
Max Pool S2	2×1		2
Convolutional S3	128×1	40	2
ReLU. S3			
Max Pool S3	2×1		2
Convolutional S4	64×1	60	2
ReLU. S4			
Max Pool S4	2×1		2
Dropout 30%			
Flatten			
Fully Connected	1×5		
Regression Output	1×5		
Total Parameters	322,955		

Early stopping is used to prevent overfitting, saving the best performing model over the validation partition. This occurred at epoch 209, with an average validation MSE (over the 5 output parameters) of 0.08572. The results obtained over the test partition and over a real AI-IoT network deployment are discussed in the results section (Sec. VI).

V. IoT System Design

In this section we will describe the design of the IoT system that performs in-node predictions of room acoustical parameters using the trained CNN model described above. The system is designed to be easily deployed in a room by conveniently distributing receiver nodes within the monitored space with the intention of obtaining a description of the acoustic environment, through the prediction of room acoustics and intelligibility parameters. The platforms and protocols used have been selected to simplify network deployment and configuration issues. Moreover, the proposed CNN-based approach allows to predict with low computational cost the considered parameters from an active speech source, without the need to carry out expensive and bulky acoustic equipment to perform RIR measurements. However, for reproducibility purposes in our experiments, we reproduce a speech signal from a loudspeaker connected to one of the nodes.

A. IoT Node Devices

The implemented IoT system consists, from an acoustic point of view, of several receiver nodes and 1 control node. All nodes are made up of a Raspberry Pi 3B board powered by a 3800 mAh battery. The receiver nodes include a low-cost lavalier microphone with USB connection (see Fig. 9 (left)). It has omnidirectional reception pattern, a sensitivity of -30 dB (+/-3 dB) and a bandwidth of 20 Hz-16 kHz. The control node is connected to a 30 Watt loudspeaker to reproduce a prerecorded anechoic speech signal during the experiments. The choice of both, the SBC board and the peripherals, was

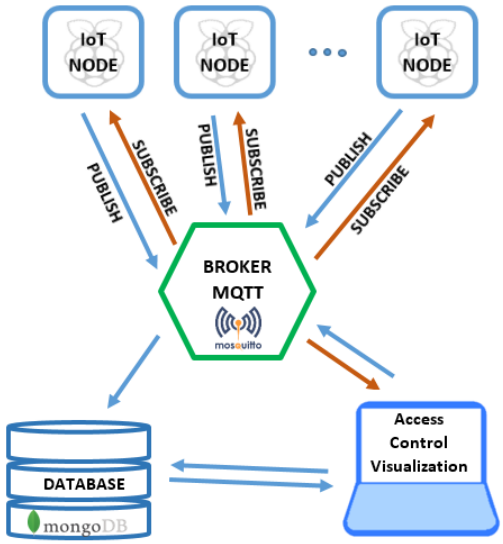


Fig. 2. Diagram of the IoT system operation.

based on a trade-off among price, processing capacity, ease of connection and low power consumption. Note also that the system makes use of speech signals, allowing us to use microphones with moderate quality and limited frequency response. These, far from ideal characteristics, do not interfere in the processes of communication, data collection and storage of the data in the sensor network, but may affect the prediction of the parameters by the CNN model, since the signals acquired by the receivers may present different characteristics to those contained in the training dataset. Thus, to analyze the potential impact of such mismatches, we will analyze as well the performance of the network over a real scenario in Sec. VI.

B. IoT System Operation

Figure 2 shows a diagram of the IoT system operation. Communication is based on the IoT protocol Message Queue Telemetry Transport (MQTT), which transmits information using messages between the nodes and the MQTT broker. In our case, the MQTT broker is implemented in the control node using Eclipse Mosquitto. In the MQTT broker, the "Room/Control", "Room/Dimension" and "Room/Nodes" topics have been created to implement the synchronization of the nodes and to collect information from each node present in the room.

Once the system starts, the MQTT broker is initialized and the "Room/Control" topic is used to init and synchronize the room parameter calculation process at regular time intervals. The transmitting nodes that are subscribed to this topic, predict the room parameters using the previously trained CNN and publish the predicted values in the "Room/Nodes/Node**/Parameters" topic, identified by the node number.

MQTT allows 3 levels of quality of service (QoS) to verify the delivery of messages and various security mechanisms regarding the transmitted data as well. In our case, we have chosen the highest quality level, QoS-2, which guarantees the delivery of messages only once, without losses or duplications. In terms of security, we have also implemented the most complete package using username and password, both in the broker and in the clients, and encryption based on SSL certification for the transmitted data. As the Nodes can create a new topic just by publishing in it, to add more nodes to the IoT system, they only have to post in the "Room/Nodes" topic, which greatly facilitates the scalability of the system.

The data received in the "Room/Nodes" topic is stored locally in a database (implemented using Mongo DB), where we can access to visualize the data. This data can also be stored in the cloud, providing a backup copy and extra security against data loss. For visualizing graphically the data, the positions of the nodes and the dimensions of the room to be analyzed must be indicated in the topics "Room/Nodes/Node**/Position" and "Room/Dimension" respectively. Otherwise, default dimensions and positions are used in the form of a rectangle with the nodes distributed equidistantly inside. This allows us not only to visualize the data in the form of a table, but also to perform a visualization in the form of heat maps, as observed in Figure 10, with the prediction of the SII parameter in one of the tested classrooms.

VI. Results

In this section we will describe the results obtained by evaluating the proposed acoustic parameter prediction framework in terms of prediction error and computational performance. We evaluate the performance of the CNN model over the independent test partition, but we also analyze the results obtained from field test, using a WASN deployed within a real classroom. To complete our evaluation, we analyze the response time using different hardware platforms, including SBC. The parameters to be predicted by the CNN, are the five mentioned previously: $RT60$, $C50$, $C80$, STI and SII. The tests described in this section are aimed at elaborating on the generalization capabilities of the system, which reflect its ability to adapt to audio signals coming from different sources, recorded with different devices and acquired in acoustic environments unseen during the development stage.

A. Prediction Accuracy

The results presented in this section will analyze the performance of the CNN model with respect to its capability to estimate the underlying reference parameters from a given input speech signal. This performance is analyzed exclusively for the trained CNN model, without taking into account any form of information fusion or averaging across node predictions. As already mentioned in Sec. (IV-B), at the end of the training process, the network achieved a MSE of 0.08572 over the validation dataset. However, in order to assess the performance over a completely independent set of data, the model is evaluated by considering the separate test partition, that includes speech signals and responses that did not participate in the training process. By doing so, we obtain a more accurate view of its real performance. The results of such evaluation are collected in Table III, where the MSE, Mean Absolute Error

TABLE III
PREDICTION MSE, MAE AND MRE EVALUATION IN INDEPENDENT TEST DATASET PARTITION

Parameter (unit)	MSE	MAE	MRE (%)	Error σ	ρ
$RT60$ (s)	0.0134	0.0652	5.59	0.0957	0.9951
$C50$ (dB)	0.2100	0.2788	7.82	0.3636	0.9969
$C80$ (dB)	0.2235	0.2921	9.29	0.3718	0.9978
STI (unitless)	0.0003	0.0132	2.19	0.0135	0.9919
SII (unitless)	0.0008	0.0205	2.28	0.0211	0.9697
$Average$	0.0896	0.1339	5.43	0.1732	0.9903

(MAE), Mean Relative Error (MRE) and Pearson correlation coefficient (ρ) for each parameter are shown. In this case, as expected, when using audios from an independent dataset, the general MSE slightly raises up to 0.0896, in the predictions of these 5 parameters, with an average correlation coefficient value greater than 0.99. Then, the final performance on the test set does not deviate significantly from the obtained over the validation data and the model seems to behave properly on unseen data.

In terms of MRE, the error for all the parameters is always below 10%, with the worst case given by $C80$ (9.29%), and close to 2% for STI and SII. Note that the accuracy achieved may be enough for common acoustic monitoring applications, as the standard deviation of errors tend to be below the just noticeable difference (JND) [47], [48].

Figures 3, 4, 5, 6 and 7 show scatter plots for the values predicted by the model against the actual ones, providing a more descriptive view of the prediction error. Each point represents a predicted value (y-axis) versus its true reference value (x-axis) for a single audio example in the dataset. Perfect predictions would lie on the diagonal dotted line. The mean MRE considering all the parameters is slightly below 5.5% as observed in Table III. This value is mainly due to $C50$ and $C80$ and to a lesser extent to $RT60$.

By analyzing Fig. 3, is evident that the $RT60$ error increases considerably (points further away from the ideal line) for values higher than 2 seconds. This is probably due to the fact that most reverberation values in the training set are between 0.1 and 1.5 seconds. In the case of the $C50$ and $C80$ parameters, the largest relative errors are obtained for values close to 0 dB, since small deviations cause a very high relative error, despite the network correctly predicts very low clarity values, as shown in Figs. 4 and 5.

There are not many studies describing CNN models aimed at predicting simultaneously more than one acoustic parameter from raw speech. Some existing approaches have been proposed to predict some of the parameters considered in this work isolatedly. In this context, while the models in [21], [22] and [23] were designed to predict $RT60$, $C50$ and STI respectively, [24] and [25] consider the estimation of multiple parameters. More concretely, [24] estimates two parameters, namely $RT60$ and direct-to-reverberant ratio (DRR), while [25] considers a more extensive set, including $RT60$, $C80$ and STI. As all these previous works have used different datasets and the trained models are not publicly available usually, it is difficult to provide direct and accurate comparisons.

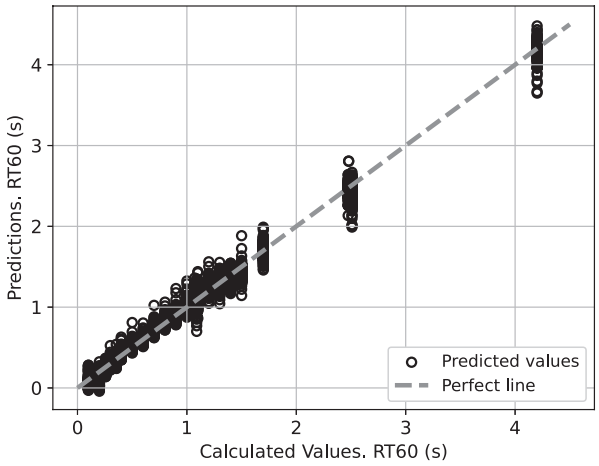


Fig. 3. True $RT60$ (ground-truth) versus predicted in the test dataset.

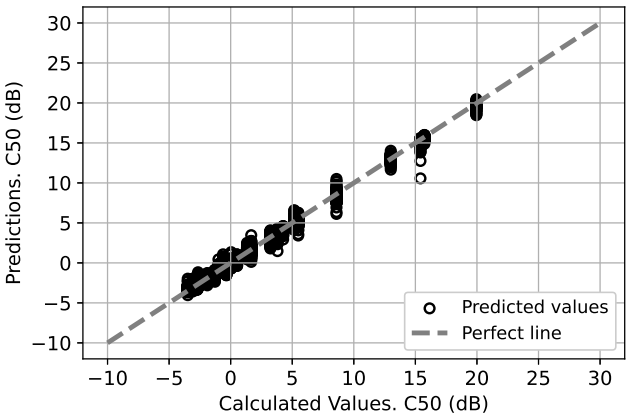


Fig. 4. True $C50$ (ground-truth) versus predicted in the test dataset.

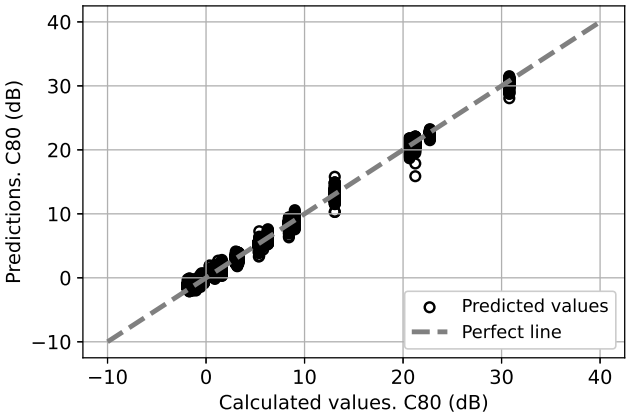


Fig. 5. True $C80$ (ground-truth) versus predicted in the test dataset.

Nonetheless, Table IV compiles the reported performance

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3203570

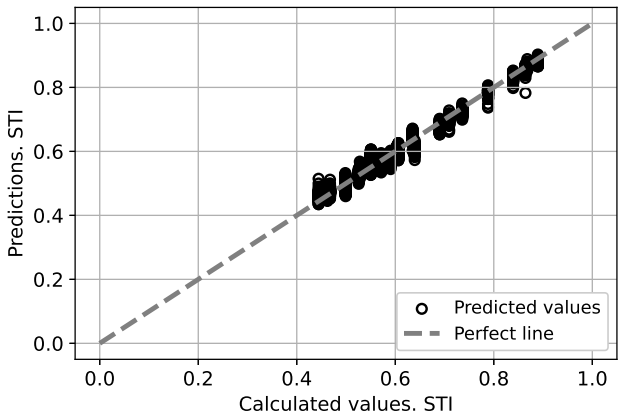


Fig. 6. True STI (ground-truth) versus predicted in the test dataset.

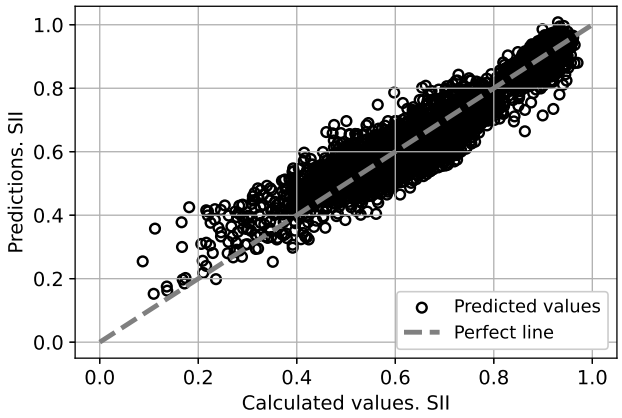


Fig. 7. True SII (ground-truth) versus predicted in the test dataset.

metrics in terms of RMSE and Pearson correlation coefficient (ρ) for those works and the ones obtained by our model in our test dataset. Although the table may be useful to confirm that the values achieved are aligned to those of the state of the art, special care should be taken when comparing such values due to the above considerations. To perform a more accurate comparison, we evaluated our model using the data from the ACE Challenge Corpus [49], used also in [21] for evaluation purposes. Table V shows the MSE results obtained for RT60 predictions on the ACE Challenge evaluation dataset. The performance obtained on this dataset without retraining our model was slightly below the one reported in [21], which also used the ACE Challenge development data in the training stage. Therefore, we also retrained the model by adding to our training dataset the ACE Challenge development data without any modification, although we needed to extract ground-truth values for the rest of parameters predicted by our model. By adding the new data, the model significantly improved the performance on the ACE Challenge evaluation dataset, confirming the effectiveness of the proposed design.

B. Computation Time Performance

In the case of room acoustic parameters, it may seem that it is not essential to be able to work in real time. However, if we want a measure of intelligibility at different positions in a room, or continuously while a speech or lecture is given, it is very important to be able to obtain these parameters as fast as possible. The computation speed is essential also in case of in-node calculation within a WASN, where the processing capacity is constrained. Thus, to perform this test, we have taken a total of 1000 audio signals from the test partition and measured the time required to calculate these parameters directly, from the impulse responses in the case of $RT60$, $C50$, $C80$ and STI and from the speech signal itself in the case of SII. Then, we measured the time spent by the CNN in predicting the same parameters, in this case only from the speech signals. To gain a better insight into the prediction times obtained on different devices, we repeated this test on 4 different platforms: 2 personal computers and 2 SBCs with the following technical specifications:

- Desktop computer: Intel(R) Core(TM) i7-7700 CPU @ 3.60 GHz, 32 GB RAM, 1 TB HDD.
- Laptop computer: Intel(R) Core(TM) i7-1065G7 CPU @ 1.3 GHz, 16 GB RAM, 500 GB SSD.
- Udoo X86 II-Ultra board: Intel(R) Pentium (TM) N3710 @ 2.56 GHZ, 8 GB RAM, 500 GB HDD.
- Raspberry PI 3B board: Broadcom BCM2837 CPU @ 1.2 GHz, 1 GB RAM, 16 GB SDHC class 10.

As we can see in Fig. 8, using a Desktop computer, the CNN model allows to predict the room acoustic parameters in an average time of 0.0039 seconds, compared to 0.9590 seconds of average time by direct calculation. This shows that the prediction using our model is 245.8 times faster than the direct signal processing calculation. On the laptop computer the ratio drops a bit to 211.3 times, however it is still a considerable advantage.

Due to the high accuracy of the CNN and the interest in bringing the model to SBC devices for AI-IoT applications, we evaluated its performance on this family of platforms as in [17], in particular using Raspberry Pi and Udoo X86, for in-node calculation. Due to the lower specifications of these devices, they present higher times in both computation and prediction, but it still demonstrates a clear advantage of the CNN-based approach over direct computation, as it is 78.8 times faster on the Udoo and up to 104.1 times faster on the Raspberry Pi. This increase in the speed of obtaining the acoustic parameters of the room represents a clear advantage over the direct calculation, taking into account that obtaining the RIR for the calculation of certain parameters is more complex and time consuming. In addition, the IoT system with the embedded AI model would allow the continuous monitoring of a room, while a class or a conference is taking place without having to interrupt it, since the parameters are directly predicted from the speech signals. The acoustic conditions of a room may vary depending on its occupation by people, or with furniture modifications, causing undesired effects in the form of reinforcement of undesired modes or absorption of certain frequency bands. A system that allows continuous monitoring

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3203570

TABLE IV
PERFORMANCE COMPARISON WITH OTHER CNN MODELS, RMSE AND CORRELATION COEFFICIENT.

	RT60		C50		C80		STI	
Model Ref.	**RMSE**	ρ	**RMSE**	ρ	**RMSE**	ρ	**RMSE**	ρ
[21] *	0.196	0.836	-	-	-	-	-	-
[22] *	-	-	3.300	0.840	-	-	-	-
[23] *	-	-	-	-	-	-	0.037	No data
[24] **	0.161	0.919	-	-	-	-	-	-
[25] **	0.393	0.918	-	-	2.038	0.943	0.040	0.913
Ours **	**0.115**	**0.995**	**0.458**	**0.996**	**0.472**	**0.997**	**0.017**	**0.991**
*	Models predicting only one parameter							
**	Models predicting more than one parameter							

TABLE V
PERFORMANCE COMPARISON WITH REF.[21] MODEL, USING ACE CHALLENGE DATASET.

	RT60	
Model Ref.	**MSE**	ρ
[21]	0.0384	0.836
Ours without retraining	**0.0421**	**0.8101**
Ours with retraining	**0.0309**	**0.9316**

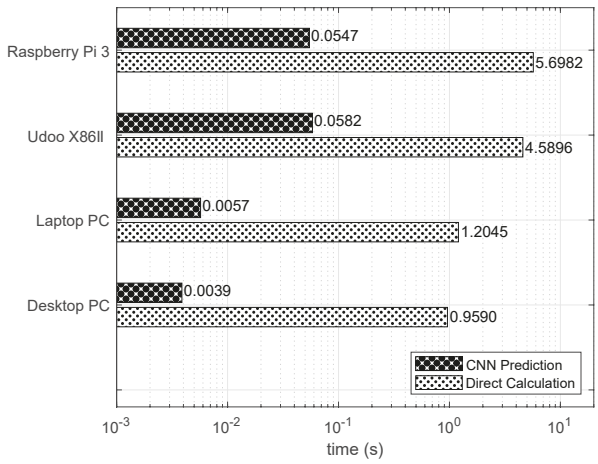


Fig. 8. Elapsed time in direct calculation and elapsed time in prediction on different platforms.

makes it possible to observe how acoustic parameters vary in real time and to act accordingly if necessary. In addition, this increase in computational speed represents a significant energy saving, allowing the IoT system to operate for hours in the absence of a connection to the power grid through the use of batteries.

C. Field Test in a Real Classroom

This test is aimed at evaluating the performance of the full AI-IoT framework described in Sec. V in some real classrooms in the facilities of the School of Engineering (ETSE) of the University of Valencia. More specifically, this test constitutes a proof-of-concept of the proposed framework, as it is intended to validate the capability of the AI model to predict with accuracy the actual value of the considered acoustical parameters. The tested system is made up of 6 receiver nodes and 1 control node, placed as depicted in Fig. 10 for the case of classroom 1.1.3. In order to compute true reference values for the different parameters, impulse response measurements were previously performed at the same positions where the nodes of the IoT system were placed (see Fig. 9 (right)). Impulse response measurements were performed according to ISO 3382 Part 1 and 2 [2], [3], using the Matlab app "Impulse Response Measurer" and the following equipment:

- Measurement microphone: Behringer ECM8000.
- Professional recording soundcard: Presonus 1818-VSL.
- Active studio monitor speaker: ESI nEar 05.

For test reproducibility purposes, speech signals different from those used in the development stage of the system were reproduced through the loudspeaker and recorded by the different nodes. Each of the nodes recorded audio with a sampling rate of 44.1 kHz, downsampling to 16 kHz and trimming the recorded signal to the input size of the model before feeding the network for predictions. The process was repeated 10 times in order to average several predictions and obtaining a more reliable measurement.

Note that the occurrence of room modes results in the reverberation time being position-dependent, since the room response at a specific location will depend on the most dominant modes at that location. It is a common practice to average multiple measurements from different locations, since a single measurement at a single point will not usually give a representative reverberation time for the room. However, since our objective was to evaluate how well each node predicts the parameters corresponding to its own location, the results compare the true reference values obtained at each node position (no averaging is performed across node predictions).

Table VI shows the calculated values and the mean of the predictions for each of the six positions, together with the standard deviation of the error, for each classroom position analyzed and for each parameter. The last row of the table shows the mean absolute error in prediction of the 6 nodes. As expected, the prediction error is somewhat higher than the one obtained when evaluating the test partition. This is probably due to the mismatch faced by the model, which is now predicting on signals captured within rooms that not only have not participated in the training process, as is the case for those of the test dataset, but which in this case may also present unwanted features such as microphone self-noise, ambient background noise or other overlapping sounds causing interference.

Nonetheless, the overall system performance returns the average MAE in prediction across the 6 nodes shown in last

TABLE VI
CNN FIELD TEST, USING A 6-NODE WASN IN A CLASSROOM.

WASN	$RT60$ (s)		$C50$ (dB)		$C80$ (dB)		STI		SII	
Node	Calc.	Avg. Pred. (σ)	Calc.	Avg. Pred. (σ)	Calc.	Avg. Pred. (σ)	Calc.	Avg. Pred. (σ)	Calc.	Avg. Pred. (σ)
1	0.85	0.84 (0.10)	3.56	2.81 (1.10)	6.53	5.63 (1.28)	0.64	0.64 (0.03)	0.74	0.66 (0.06)
2	0.84	0.80 (0.08)	3.04	2.56 (1.05)	6.18	5.39 (1.24)	0.65	0.64 (0.02)	0.73	0.64 (0.06)
3	0.77	0.91 (0.12)	3.52	3.28 (1.18)	7.06	6.25 (1.40)	0.65	0.65 (0.02)	0.73	0.71 (0.08)
4	0.81	0.86 (0.12)	3.27	3.64 (1.02)	6.85	6.60 (1.19)	0.66	0.66 (0.02)	0.75	0.72 (0.06)
5	0.81	0.82 (0.10)	3.28	3.27 (1.03)	6.15	6.15 (1.24)	0.66	0.65 (0.02)	0.76	0.71 (0.06)
6	0.79	0.83 (0.11)	4.30	4.24 (1.05)	7.44	7.28 (1.25)	0.67	0.67 (0.02)	0.77	0.77 (0.06)
MAE		0.0483		0.3183		0.4850		0.0033		0.0450

row of Table VI, with values below the JND values [47], [48] although with less distance than in the test partition evaluation. The prediction errors for $C50$, $C80$ and SII are little higher and may be due to the unwanted effects discussed above, however the errors for $RT60$ and STI are slightly lower than in the test partition evaluation.

In any case, the predictions obtained in this test were reasonably accurate for our application, computationally efficient and free of measurement procedures involving the playback of specific test signals. This confirms the validity of the proposed framework for AI-IoT monitoring of room acoustic and intelligibility parameters, allowing to obtain a meaningful description of the acoustic behavior of a room in a very convenient way. As an example, Fig. 10 shows a heat map generated from the SII predictions that indicates the level of speech intelligibility as a function of the position in the classroom. Simulating a common location of the lecturer, the blue circle indicates the loudspeaker source position and the arrow its orientation. The red circles indicate the node positions which are uniformly distributed within the classroom. It can be clearly observed that speech intelligibility decreases with the distance to the source, which is the expected behavior. Note that this framework can bring significant advantages at a very low cost to monitor spaces with conflicting areas where speech perception is poor, constituting a simple, accurate and fast method. In spite of the successful results obtained above, the proposed system may face some limitations. The architecture considers a one-channel input of 56000 samples with a sampling rate of 16 kHz, limiting the analysis to a maximum frequency of 8 kHz. Since the spectral content of speech is mostly contained within the considered range, this may not be problematic, but there is a clear limitation to cover higher octave bands. In the time domain, the input signal has a duration of 3.5 seconds, limiting the maximum reverberation time that may be accurately analyzed. This aspect, which is not relevant in the analysis of conventional rooms, poses limitations when studying very large enclosures with high reverberation times, such as a cathedral for example. Nevertheless, it is worth mentioning that retraining the system using signals with other spectro-temporal characteristics would mitigate these effects.

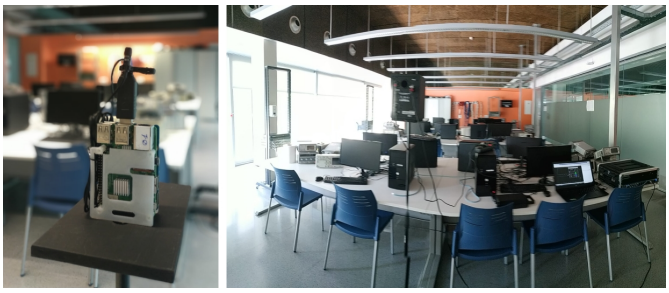


Fig. 9. IoT Node (left). Classroom 113 while performing RIRs measurement process (right).

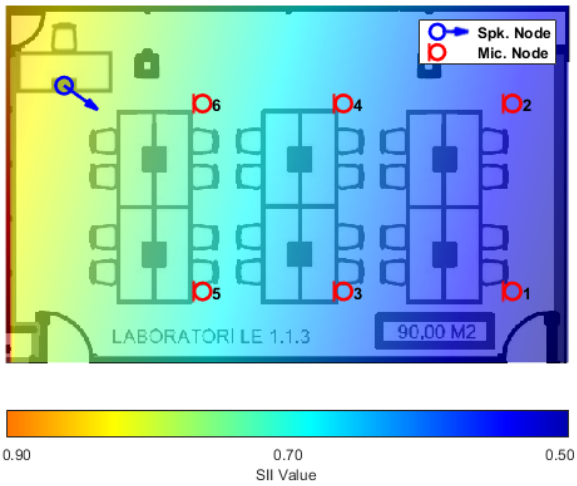


Fig. 10. Speech Intelligibility Index (SII) prediction on the field test, Classroom 113.

VII. CONCLUSION

In this work, we have introduced and developed a full AI-IoT framework for estimating blindly a set of room acoustic and intelligibility parameters ($RT60$, $C50$, $C80$, STI and SII) directly from speech signals. The proposed system is based on a lightweight convolutional Neural Network (CNN) model trained on a dataset of speech signals convolved with synthetic and real room impulse responses with varied acoustic properties. The resulting model achieved an average mean relative error of less than 5.5% on the test partition and demonstrated to be 245 times faster than the direct signal processing calculation. Additionally, the model has been integrated into an IoT-based communication scheme, allowing continuous and simultaneous monitoring of the acoustic parameters at different node locations without the need to perform cumbersome acoustic measurements. Thus, the full system offers an accurate, flexible and computationally efficient solution for acoustic monitoring applications at a reduced cost.

ACKNOWLEDGMENT

Authors would like to thank the Spanish State Research Agency (AEI) and the European Regional Development Fund (ERDF) for partially funding this research within the projects with grant references PID2021-126823OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR", RTI2018-097045-B-C21 funded by MCIN/AEI/ 10.13039/501100011033 and by "ERDF A way of making Europe" and the grant BES-2017-082340 funded by MCIN/AEI/ 10.13039/501100011033 and by "ESF Investing in your future". Also to the Generalitat Valenciana, for funding the grant AEST/2020/048, AEST/2021/16, and projects GV/2020/046 and AICO/2020/154. Finally, to the Universitat de València for funding the special action UV-INV-AE-1544281.

REFERENCES

[1] S. Weinzierl and M. Vorländer, "Room acoustical parameters as predictors of room acoustical impression: What do we know and what would we like to know?" *Acoust Aust*, vol. 43, p. 41–48, 2015. [Online]. Available: https://doi.org/10.1007/s40857-015-0007-6

[2] ISO, "ISO 3382-1:2009. Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces." UNE-EN, Tech. Rep., Mar. 2016.

[3] ——, "ISO 3382-2:2008. Acoustics – Measurement of room acoustic parameters – Part 2: Reverberation time in ordinary rooms." UNE-EN, Tech. Rep., Dec. 2008.

[4] ——, "ISO 3382-3:2012. Acoustics – Measurement of room acoustic parameters – Part 3: Open plan offices." UNE-EN, Tech. Rep., Oct. 2012.

[5] ——, "ISO 9921: 2004, Ergonomics - Assessment of speech communication," UNE-EN, Tech. Rep., Jun. 2008.

[6] ANSI/ASA, "ANSI/ASA S3.5-1997 (R2017), American National Standards Institute. Acoustical Society of America. Methods For Calculation Of The Speech Intelligibility Index," Tech. Rep., Jun. 1997.

[7] Y. Ando, *Concert Hall Acoustics*, 1st ed. Berlin, Heidelberg: Springer-Verlag, 7 1985, vol. 17.

[8] L. Beranek, *Concert Halls and Opera Houses*, 2nd ed. New York: Springer-Verlag, 7 2004.

[9] Odeon A/S, "ODEON Room Acoustics Software. User's Manual. Version 17," 2021, accessed: 17/05/2022. [Online]. Available: https://odeon.dk/download/Version17/OdeonManual.pdf

[10] B&K, "Measuring Speech Intelligibility using DIRAC Type 7841," 2013, accessed: 17/05/2022. [Online]. Available: https://www.bksv.com/-/media/literature/Application-Note/bo0521.ashx

[11] "Raspberry Pi 3B+," 2018, accessed: 02/12/2020. [Online]. Available: https://www.raspberrypi.org/

[12] A. Alexandridis and A. Mouchtaris, "Multiple sound location estimation and counting in a wireless acoustic sensor network." in *Proc of 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. October 18-21, 2015, New Paltz, NY*, 2015.

[13] B. Malhotra, I. Nikolaidis, and J. Harms, "Distributed classification of acoustic targets in wireless audio-sensor networks," *Computer Networks*, 2008.

[14] M. F. Duarte and Y. Hen Hu, "Vehicle classification in distributed sensor networks." *J. Parallel Distrib.Comput.*, vol. 64, p. 826838, 2004.

[15] A. Pastor-Aparicio, J. Segura-Garcia, J. Lopez-Ballester, S. Felici-Castell, M. Garcia-Pineda, and P.-S. J. J., "Psychoacoustic annoyance implementation with wireless acoustic sensor networks for monitoring in smart cities," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 128–136, 2020.

[16] J. Lopez-Ballester, J. Alcaraz Calero, J. Segura-Garcia, S. Felici-Castell, M. Garcia-Pineda, and M. Cobos, "Speech intelligibility analysis and approximation to room parameters through the internet of things," *Applied Sciences*, vol. 11, no. 4, p. 1430, 2021.

[17] J. Lopez-Ballester, A. Pastor-Aparicio, S. Felici-Castell, J. Segura-Garcia, and M. Cobos, "Enabling real-time computation of psychoacoustic parameters in acoustic sensors using convolutional neural networks," *IEEE Sensors Journal*, vol. 20, no. 19, pp. 11 429–11 438, 2020.

[18] A. F. Genovese, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Blind room volume estimation from single-channel noisy speech," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 231–235.

[19] W. Yu and W. B. Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2020.

[20] C. Papayiannis, C. Evers, and P. A. Naylor, "End-to-end classification of reverberant rooms using DNNs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3010–3017, 2020.

[21] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 136–140.

[22] P. Peso Parada, D. Sharma, J. Lainez, D. Barreda, T. v. Waterschoot, and P. A. Naylor, "A single-channel non-intrusive c50 estimator correlated with speech recognition performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 719–732, 2016.

[23] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind estimation of the speech transmission index for speech quality prediction," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 591–595.

[24] N. Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," 05 2020, pp. 1–5.

[25] S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response," *Applied Acoustics*, vol. 185, p. 108372, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X21004667

[26] C. L. C. Lam, "Improving the speech intelligibility in classrooms," Ph.D. dissertation, Dept. of Mechanical Engineering, The Hong Kong Polytechnic University, The address of the publisher, 2010. [Online]. Available: https://theses.lib.polyu.edu.hk/handle/200/6621

[27] R. McNeer, C. Bennett, D. Horn, and D. R., "Factors affecting acoustics and speech intelligibility in the operating room: Size matters." *Anesth Analg.*, vol. 124, no. 6, pp. 1978–1985, 2017.

[28] H. T. Ryherd EE, Moeller M Jr, "Speech intelligibility in hospitals." *J Acoust Soc Am.*, vol. 134, no. 1, pp. 586–595, 2013.

[29] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *AES 108th Convention, Paris February 19-22*, 11 2000.

[30] M. Vorlander and M. Kob, "Practical aspects of MLS measurements in building acoustics," *Applied Acoustics*, vol. 52, no. 3–4, pp. 239–258, 1997.

[31] G.-B. Stan, J.-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249–262, 2002.

[32] L. G. Marshall, "An acoustics measurement program for evaluating auditoriums based on the early/late sound energy ratio," *The Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2251–2261, 1994. [Online]. Available: https://doi.org/10.1121/1.410097

[33] D. Gottlob, *Vergleich objektiver akustischer Parameter mit Ergebnissen subjektiver Untersuchungen an Konzertsälen*. Georg August Universität zu Göttingen., 1973.

[34] M. R. Schroeder, D. Gottlob, and K. Siebrasse, "Comparative study of european concert halls: correlation of subjective preference with geometric and acoustic parameters," *The Journal of the Acoustical Society of America*, vol. 56, no. 4, pp. 1195–1201, 1974.

[35] T. Yamamoto and F. Suzuki, "Multivariate analysis of subjective measures for sound in rooms and the physical values of room acoustics," *J. Acoust. Soc. Jpn*, vol. 32, no. 10, pp. 599–605, 1976.

[36] S. Cerdá, A. Giménez, J. Romero, R. Cibrian, and J. Miralles, "Room acoustical parameters: A factor analysis approach," *Applied Acoustics*, pp. 97–109, 01 2009.

[37] T. Houtgast and H. J. M. Steeneken, "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985. [Online]. Available: https://doi.org/10.1121/1.392224

[38] P. W. Barnett, "Review of speech intelligibility indicators: Their relationship and applications," *The Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 3050–3050, 1997.

[39] P. W. Barnett and A. M. Acoustics, "Overview of Speech Intelligibility." *Proceedings - Institute of Acoustics*, vol. 21, pp. 1–16, 1999.

[40] "American national standards institute. acoustical society of america. standards secretariat. american national standard: Methods for calculation of the speech intelligibility index." 1998, accessed: 02/12/2020.

[41] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. USA: Curran Associates Inc., 2016, pp. 892–900. [Online]. Available: http://dl.acm.org/citation.cfm?id=3157096.3157196
[42] S. G. McGovern, "Fast image method for impulse response calculations of box-shaped rooms," *Applied Acoustics*, vol. 70, no. 1, pp. 182–189, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X08000455
[43] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979. [Online]. Available: https://doi.org/10.1121/1.382599
[44] D. Murphy and D. o. E. U. o. Y. Joe Rees-Jones, Audiolab. (2021) Open acoustic impulse response (open air). [Online]. Available: https://www.openairlib.net/
[45] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.
[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
[47] J. Bradley, R. Reich, and S. Norcross, "A just noticeable difference in c50 for speech," *Applied Acoustics*, vol. 58, no. 2, pp. 99–108, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X98000759
[48] F. Martellotta, "The just noticeable difference of center time and clarity index in large reverberant spaces," *The Journal of the Acoustical Society of America*, vol. 128, pp. 654–63, 08 2010.
[49] J. Eaton, N. Gaubitch, A. Moore, and P. Naylor, "The ace challenge - corpus description and performance evaluation," 10 2015.

Jesus Lopez-Ballester received the B.Sc. and M.Sc. degrees in telecommunications engineering from the University of Valencia, Valencia, Spain, in 2009 and 2014, respectively, where he is currently pursuing the Ph.D. degree in information technology, communications and computation. He was with the Institute of Robotics, University of Valencia developing advanced machinery and vehicle simulators. His current research interests include analysis of acoustic events, e-Health, human–machine interfaces, deep learning, and signal processing. Dr. Lopez-Ballester was a recipient of the Extraordinary Prize Final Project of the College of Telecommunications Engineers for the design of a telemedicine system focused on cardiology emergencies.

Santiago Felici-Castell received the M.Sc. and Ph.D. degrees in telecommunication engineering from the Polytechnical University of Valencia, Valencia, Spain, in 1993 and 1998, respectively. He is currently an Associate Professor with the University of Valencia. He is also a Cisco Systems Certificated Instructor, and has authored over 25 technical papers in international journals and conferences. His current research interests include networking, communication systems, and multiresolution techniques for data transmission with quality of service.

Jaume Segura-Garcia received the M.Sc. and Ph.D. degrees in physics from the University of Valencia, Valencia, Spain, in 1998 and 2003, respectively. After completing his Ph.D. study, he was with the Robotics Institute, University of Valencia, where he was involved in several projects related to intelligent transportation systems. Since 2008, he has been with the Department of Computer Science, University of Valencia, where he is currently an Associate Professor. He has been a Visiting Researcher with multiple European research centers. He has coauthored over 90 publications at national and international journals, book chapters, and conferences. He was on the Organizing Committee of several national and international conferences, including the International Workshop on Virtual Acoustics (2011). He is a member of the Spanish Acoustics Society and the European Acoustics Association.

Maximo Cobos (Senior Member, IEEE) received the master's degree in telecommunications and the Ph.D. degree in telecommunications engineering from the Polytechnical University of Valencia, Valencia, Spain, in 2007 and 2009, respectively. He completed his studies with honors under the University Faculty Training Program, and was a recipient of the Ericsson Best Ph.D. Thesis Award on Multimedia Environments from the Spanish National Telecommunications Engineering Association. He received a Campus de Excelencia Postdoctoral Fellowship to work with the Institute of Telecommunications and Multimedia Applications, Valencia, in 2010. In 2009 and 2011, he was a Visiting Researcher with Deutsche Telekom Laboratories, Berlin, Germany, where he worked on the field of signal processing for spatial audio. Since 2021, he has been an Full Professor with the University of Valencia. His work is focused on the area of digital signal processing for audio and multimedia applications, where he has authored over 70 technical papers in international journals and conferences. He is a Full Member of the Acoustical Society of America.