TECNIACUSTICA ELCHE · 2022

53º CONGRESO ESPAÑOL DE ACÚSTICA
XII CONGRESO IBÉRICO DE ACÚSTICA

PLATAFORMA AI-IOT BASADA EN APRENDIZAJE PROFUNDO PARA LA ESTIMACIÓN CIEGA DE PARÁMETROS ACÚSTICOS DE SALA

PACS: 43.55.Ev, 43.55.Ka, 43.55.Br

Lopez-Ballester, Jesus; Segura-Garcia, Jaume; Felici-Castell, Santiago; Cobos-Serrano, Máximo
Departamento Informática, ETSE, Universitat de València
Avda. de la Universidad S/N, 46100, Burjassot, Valencia, España, jesus.lopez-ballester@uv.es

Palabras Clave: Internet of Things, parámetros acústicos, inteligibilidad del habla, redes neuronales convolucionales, redes de sensores.

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.

RESUMEN.
Los parámetros acústicos de sala se utilizan ampliamente para describir el comportamiento del sonido en entornos interiores, como salas de conciertos, salas de conferencias, etc. Muchos de ellos han sido estandarizados y en general su cálculo suele tener un alto coste computacional. Con la creciente presencia de enfoques basados en aprendizaje profundo en sistemas de monitorización automática, las redes de sensores acústicos inalámbricos (WASN) ofrecen un gran potencial para facilitar la estimación de dichos parámetros acústicos. En este escenario, las redes neuronales convolucionales (CNN) ofrecen reducciones significativas en los requisitos computacionales de cálculo, habilitando la predicción de parámetros en los nodos, lo que permite la llamada Inteligencia Artificial-Internet de las Cosas (AI-IoT). En este artículo describimos el diseño y análisis de un modelo de CNN basado en aprendizaje profundo, entrenado para predecir simultáneamente un conjunto de parámetros acústicos de sala conocidos, directamente a partir de señales de voz, sin necesidad de mediciones específicas de la respuesta al impulso. Los resultados muestran que la predicción de los parámetros acústicos de sala y de inteligibilidad del habla, mediante la CNN diseñada, alcanza una tasa de error relativa inferior al 5,5 %, acompañada de un incremento en la velocidad de cálculo cercano a 250 con respecto al enfoque convencional basado en procesado de señales.

1. INTRODUCCIÓN

La predicción de la impresión acústica de los recintos es un aspecto de gran interés en el diseño arquitectónico [1]. Disponer de una descripción acústica precisa o un control de la percepción sonora es especialmente relevante cuando el espacio analizado, ya sea interior o exterior, se destina a aplicaciones en las que la percepción de las fuentes musicales o del habla desempeña

un papel fundamental. En la práctica, se utilizan parámetros objetivos, comúnmente llamados parámetros acústicos de la sala (por ejemplo, el tiempo de reverberación, la claridad o la definición). Además de estos parámetros, en entornos destinados a la transmisión oral de información, se emplean descriptores que evalúan la cantidad de información oral transmitida y útil. Podemos encontrar la estandarización de estos parámetros en la ISO 3382 y sus diferentes revisiones [2] para los parámetros acústicos relacionados con reverberación y distribución energética y también la ISO 9921 y ANSI S3.5-1997 [3,4] para los relacionados con la inteligibilidad del habla. Pese a que en el diseño arquitectónico se utilizan técnicas basadas en modelos acústicos virtuales para diseñar un comportamiento acústico controlado [5], este debe ser siempre verificado mediante mediciones in situ, ya que cualquier modificación de los materiales o la geometría puede producir cambios significativos. Por ello, es necesario disponer de un sistema que permita evaluar los parámetros acústicos de la sala de forma rápida y sencilla. Existen diferentes productos comerciales que realizan el estudio y análisis de los parámetros acústicos de la sala o de la inteligibilidad del habla, pero suelen ser caros y a veces están limitados a unos parámetros específicos. En este contexto, el uso de las redes de sensores acústicos inalámbricos (WASN) puede aportar un valor añadido importante, gracias a que pueden recoger información y realizar mediciones de forma distribuida, adaptarse a diferentes tipos de análisis o actualizarse incorporando nuevos parámetros. Estas redes se han empleado en otros estudios para diferentes tareas, por ejemplo, para localizar fuentes de sonido [6], identificarlas [7] o medir características específicas del entorno [8], por citar algunas. Las redes WASN se componen habitualmente de varios nodos dotados de unidades de proceso, memoria, interfaces de comunicación y conexión, incluidos en los denominados Single Board Computers (SBC) genéricos. Debido al rendimiento generalmente moderado de estos dispositivos, las complejas tareas de procesado de señal necesarias para el cálculo de los parámetros acústicos, además de la velocidad de cálculo y capacidad de almacenamiento, es necesario idear implementaciones alternativas para realizar los cálculos. En este escenario, las redes neuronales, y en particular las redes neuronales convolucionales (CNN), ofrecen reducciones significativas en los requisitos computacionales para la predicción de parámetros.

En este trabajo, proponemos el uso combinado del aprendizaje profundo y la tecnología IoT que permite la predicción dinámica y rápida de los parámetros acústicos con un coste computacional reducido, planteando una doble contribución. Por un lado, se consigue la predicción simultánea de diferentes parámetros acústicos mediante un solo modelo de CNN, a partir de una señal de voz en crudo. Por otro lado, se realizan las predicciones in situ dentro de los nodos, sin necesidad de transmitir las señales de audio, ni del cálculo de la respuesta impulsiva de la sala. Esto ha permitido llevar el modelo a los nodos de un sistema IoT, conformando el sistema AI-IoT que se describe en este artículo, desde la generación de los datos necesarios, diseño y entrenamiento del modelo de red neuronal hasta un despliegue final del sistema en un entorno real.

2. TRABAJOS RELACIONADOS

Otros trabajos recientes han explorado también metodologías para reducir el coste computacional de los algoritmos implementados para ser ejecutados dentro de nodos IoT en aplicaciones WASN. En [9], los autores presentan resultados de un modelo de CNN aplicado al cálculo de la molestia subjetiva, basados en la estimación de un conjunto de parámetros psicoacústicos, todos ellos estimados por un modelo entrenado mediante aprendizaje profundo y con una probabilidad de error inferior al 3 %, permitiendo realizar los cálculos de forma local en los nodos. Además, podemos encontrar diferentes diseños de CNN aplicados a la estimación de diferentes características de las salas.

En [8], se propone una CNN para estimar la geometría de una sala y los coeficientes de reflexión a partir de una RI y en [10] se emplea para clasificar la sala en que se han grabado las secuencias de audio de entrada. Centrándonos estrictamente en parámetros acústicos, los trabajos presentados en [11, 12] describen modelos CNN capaces de predecir el tiempo de reverberación (RT60) y la claridad del habla (C50) respectivamente. Los parámetros acústicos de sala también son interesantes para analizar los espacios interiores con finalidades muy específicas donde la transmisión de información es vital, como por ejemplo los espacios de enseñanza o las zonas de embarque en aeropuertos o también por ejemplo del ámbito sanitario como quirófanos o salas de cirugía. Estos parámetros pueden ser empleados para evaluar el comportamiento acústico de los mencionados espacios y la adecuación a su uso [13,14].

Los trabajos anteriores confirman que la estimación de los parámetros acústicos de sala y la predicción de la inteligibilidad del habla son problemas clásicos que han atraído el interés de la comunidad investigadora durante mucho tiempo, y que siguen demandando soluciones innovadoras y desafiantes para la era de la IA-IoT, como la que se propone en este trabajo.

3. PARÁMETROS ACÚSTICOS DE SALA E INTELIGIBILIDAD DEL HABLA

La mayoría de los parámetros acústicos de salas están estandarizados en normativa como la ISO 3382 y sus diferentes revisiones [2]. La norma también hace algunas recomendaciones y especifica procedimientos sobre diferentes aspectos relacionados con el proceso de medición, pero deja algunos aspectos no vinculantes para permitir la innovación.
Diferentes investigadores seleccionan unos parámetros acústicos u otros en función de la orientación del estudio. Basándonos en diferentes estudios como los realizados por la escuela de Gottingen o Yamamoto [15] y en los estudios realizados por S. Cerdá y A. Gimenez [16], hemos seleccionado un conjunto de 5 parámetros de 3 naturalezas significativas independientes, distribución energética, reverberación e inteligibilidad del habla. Inicialmente, seleccionamos el tiempo de reverberación (RT60) y la claridad musical (C80) como parámetros representativos de los conjuntos centrados en la reverberación y la distribución energética. Añadimos también el índice de transmisión del habla (STI) [3] para tener un descriptor de inteligibilidad calculado a partir de la respuesta impulsiva de la sala que completara el conjunto. Posteriormente, razones prácticas derivadas del COVID-19 nos obligaron a realizar mediciones acústicas centradas en la inteligibilidad del habla en salas habilitadas para actividades docentes que no estaban diseñadas para ello inicialmente, por lo que incluimos la claridad del habla (C50) y el índice de inteligibilidad del habla (SII) [4] para orientar nuestro conjunto de parámetros a la evaluación de la transmisión de la información oral. Según los aspectos de la impresión subjetiva que estudian, los parámetros seleccionados para este trabajo pueden agruparse en:

1. Parámetros de reverberación, que representan el grado de viveza de una sala y se calculan a partir de la respuesta impulsiva. RT60 en concreto mide el tiempo que tarda la energía sonora en decaer 60 dB. En la práctica, estas mediciones suelen abarcar la gama de frecuencias de 50 Hz a 8 kHz en bandas de 1 octava o de 1/3 de octava.

2. Parámetros energéticos, que describen la relación entre la energía presente en las primeras reflexiones sonoras frente a la energía de las reflexiones tardías, calculándose también a partir de la respuesta impulsiva.
    a. La claridad del habla o C50, mide la proporción de energía presente en los primeros 50 ms frente a la presente desde ese punto al final de la respuesta impulsiva temporal de la sala.
    b. La claridad musical o C80, mide la proporción de energía presente en los primeros 80 ms frente a la presente desde ese punto al final de la respuesta impulsiva temporal de la sala.

3. Parámetros de inteligibilidad, que miden la proporción de habla que se puede transmitir por un canal y lo comprensible que es en determinadas condiciones:
    a. STI (Speech Transmission Index), se calcula a partir de la respuesta impulsiva, como la suma ponderada del Índice de Transferencia de Modulación (MTI), uno por cada banda de frecuencia de octava de 125 Hz a 8 kHz, donde cada valor MTI se obtiene a partir de los valores de la Función de Transferencia de Modulación (MTF) sobre 14 frecuencias de modulación diferentes, teniendo en cuenta los umbrales auditivos según la norma IEC 60268-16 o ISO 9921 [3]. Los valores de STI van de 0 que representa una muy mala transmisión del habla a 1, que representaría una transmisión excelente. En esta escala, un STI de al menos 0,5 es deseable para la mayoría de las aplicaciones.
    b. SII (Speech Intelligibility Index), que se calcula a partir de una señal de habla y se basa en la suma ponderada (por banda de frecuencia y factor de distorsión) de los niveles espectrales equivalentes de la señal del habla, y de los niveles espectrales equivalentes del ruido por banda. En nuestro caso, para su cálculo hemos utilizado la escala más precisa contemplada por la norma, en 1/3 de octava, entre 160 Hz y 8 kHz (18 bandas).

Como en el caso del STI, obtendremos un valor en el rango de 0 (muy mala) a 1 (muy buena inteligibilidad).

Como se ha descrito, todos los parámetros mencionados se calculan a partir de la respuesta impulsiva de la sala, excepto SII [4], que se puede obtener a partir de una señal de voz grabada en la posición a estudiar de la sala. El estudio del SII nos proporciona una mejor idea de cómo la posición de la fuente, la distancia, la geometría y los materiales de la sala afectan a la inteligibilidad del habla. El marco propuesto basado en IA realizará la estimación de todos los parámetros a partir de las señales de voz mediante un único modelo de CNN. Al realizar la predicción de los parámetros a partir de grabaciones reales de habla dentro de una sala, se evita la necesidad de engorrosas mediciones de la RI.

4. REDES NEURONALES CONVOLUCIONALES PARA PREDICCIÓN DE PARÁMETROS

Las CNNs han demostrado ser muy útiles para predecir parámetros relacionados con la acústica de forma rápida y precisa [9], debido principalmente a su capacidad para aprender representaciones optimizadas de la señal durante el proceso de entrenamiento. Podemos definir esquemáticamente una red neuronal como un conjunto de capas formadas por neuronas. Con un entrenamiento adecuado, la combinación de capas puede aprender un mapeo de las entradas a las salidas proporcionadas en un conjunto de entrenamiento. Las CNNs se caracterizan por tener capas convolucionales que implementan diferentes procesos de filtrado. Los filtros específicos se definen durante el proceso de entrenamiento ajustando los pesos de la red, minimizando el valor de una función de pérdida definida. En una red de extremo a extremo como la que se propone en este trabajo, la entrada al modelo es un segmento de audio en bruto, y es la red la que debe analizar y extraer características significativas a lo largo de sus sucesivas etapas. A pesar de que las CNNs se conocen por ser aplicadas a tareas de clasificación de señales, especialmente en el dominio de la imagen, los parámetros acústicos mencionados varían a lo largo de una escala continua, por lo tanto, abordamos el diseño de nuestra CNN orientada a la predicción de parámetros como un problema de regresión. Así, el objetivo es obtener una arquitectura precisa de extremo a extremo capaz de predecir todo el conjunto de parámetros a partir de entradas de audio en bruto.

4.1. Conjunto de datos

El proceso de entrenamiento de una red neuronal requiere una gran cantidad de datos de audio acompañados de las etiquetas correspondientes en el caso de que se trate de clases y que en nuestro caso corresponden a los valores reales de los 5 parámetros considerados. Como encontrar gran número de salas con características diferentes es complejo, se ha generado sintéticamente un conjunto discreto de respuestas al impulso correspondientes a 15 salas rectangulares diferentes con valores de RT60 comprendidos entre 0,1 y 1,5 segundos, mediante el método imagen-fuente, variando dimensiones y coeficientes de reflexión en los diseños de salas. Para cada sala sintética se generó un conjunto de 20 RIs según diferentes configuraciones fuente-micrófono considerando una retícula rectangular. Para completar el conjunto de respuestas de impulso, incluimos respuestas de otras 10 salas reales extraídas del repositorio OpenAir (Open Acoustic Impulse Response Library). Así, se han utilizado un total de 310 RIRs procedentes de 25 salas para crear un conjunto de datos de señales de habla grabadas en diferentes entornos acústicos. Las muestras de habla utilizadas para generar el conjunto de datos se extrajeron de la base de datos acústico-fonética de habla continua DARPA TIMIT. Los pasos de pre-procesamiento aplicados a las señales de voz extraídas han sido: remuestreo a 16 kHz para tener tasas uniformes, convolución con las diferentes RI y recorte del resultado en secuencias de 3,5 segundos (56000 muestras). Para realizar el entrenamiento y la evaluación de la red neuronal, el conjunto de datos completo se dividió en dos particiones de datos: entrenamiento y validación, añadiendo una partición de prueba posteriormente. Las particiones que participan en el proceso de entrenamiento son únicamente la de entrenamiento y la de validación. Para estos dos conjuntos de datos, se han empleado respectivamente el 80% para entrenamiento (26000 señales) y el 20% para la validación (4000 señales), del total de 30000 empleadas. Además, hemos añadido un conjunto de 1000 señales que forman la partición de

prueba, utilizada para probar el rendimiento de la CNN más a fondo con señales que no hayan participado en el proceso de entrenamiento. Esta última partición de prueba está compuesta por señales de voz no empleadas en los otros conjuntos, convolucionadas con 15 respuestas impulsivas que tampoco se han empleado anteriormente.

4.2. Diseño, configuración y entrenamiento

El diseño de la CNN se basa en nuestro trabajo anterior [9], donde entrenamos una CNN para predecir parámetros de molestia psicoacústica. Esquemáticamente consta de varias capas convolucionales seguidas de capas *Max Pool,* donde el tamaño de los filtros va disminuyendo, mientras que su número aumentará a medida que la red se hace más profunda.
La arquitectura empleada que se puede ver con detalle en la Tabla 1, se basa en 4 etapas convolucionales (S1 a S4), formadas cada una por una capa de convolución temporal, una capa de activación *Rectified Linear Unit* (ReLU), que elimina valores negativos y una capa Max Pool, que conserva únicamente el máximo de un conjunto de elementos, implementando un submuestreo de la entrada de la capa. Seguido a 4 unidades como la descrita, se sitúa una capa *Dropout* con una probabilidad de abandono de 0,3 para evitar el sobreajuste, seguida de una capa *Flatten* que convierte la salida multidimensional en un vector de características unidimensional. Por último, una capa *Fully connected* es seguida por una capa *Output Regression* destinada a minimizar el MSE de los 5 parámetros considerados: RT60, C50, C80, STI y SII.

Tabla 1 – Definición de capas

Capa	Tamaño	Nº Filtros	Paso
Input	56000_1		
Convolutional S1	512x1	10	10
ReLU S1			
Max Pool S1	2x1		2
Convolutional S2	256x1	20	5
ReLU S2			
Max Pool S2	2x1		2
Convolutional S3	128x1	40	2
ReLU S3			
Max Pool S3	2x1		2
Convolutional S4	64x1	60	2
ReLU S4			
Max Pool S4	2x1		2
Dropout 30%			
Flatten			
Fully Connected	1x5		
Regression Output	1x5		

El entrenamiento se realizó utilizando un optimizador Adam con una tasa de aprendizaje inicial de 10^{-3} y empleando lotes de 512 señales barajadas en cada iteración. La función de pérdida que debe minimizarse durante el proceso de entrenamiento es el error cuadrático medio (MSE), que se obtiene de las diferencias cuadráticas medias entre los valores verdaderos y los predichos. El MSE es muy sensible a los valores atípicos que difieren de la media, por lo que se ajusta bien a los problemas basados en la regresión donde se pretende penalizar más a los errores grandes frente a los pequeños. 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. El MSE sobre el conjunto de datos de validación, obtenido al final del proceso de entrenamiento es de 0,08572 (sobre los 5 parámetros de salida).

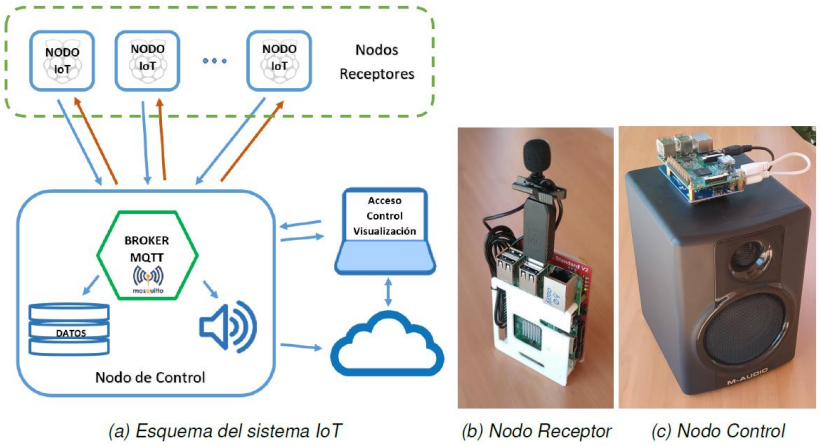
5. SISTEMA IOT

En esta sección describiremos el diseño del sistema IoT que realiza predicciones en el nodo de los parámetros acústicos de la sala utilizando el modelo CNN entrenado descrito anteriormente. El sistema está diseñado para ser fácilmente desplegado en una sala distribuyendo

convenientemente los nodos receptores dentro del espacio monitorizado. El enfoque propuesto basado en CNN permite predecir con un bajo coste computacional los parámetros considerados a partir de una fuente de habla activa, sin necesidad de llevar a cabo costosos y voluminosos equipos acústicos para realizar mediciones de RI. Sin embargo, a efectos de reproducibilidad en nuestros experimentos, reproducimos una señal de habla procedente de un altavoz conectado a uno de los nodos.

5.1. Nodos IoT

El sistema IoT implementado consta, desde el punto de vista acústico, de varios nodos receptores y 1 nodo de control, como se puede ver en la Figura 1a. Todos los nodos están formados por una placa Raspberry Pi 3B alimentada por una batería de 3800 mAh. Los nodos receptores incluyen un micrófono de solapa de bajo coste con conexión USB (véase la Fig. 1b). Tiene un patrón de recepción omnidireccional, una sensibilidad de -30 dB (+/-3 dB) y un ancho de banda de 20 Hz-16 kHz. El nodo de control está conectado a un altavoz de 30 vatios para reproducir una señal de habla anecóica (véase la Fig. 1c). La elección de ambos, la placa SBC y los periféricos, se basó en un compromiso entre el precio, la capacidad de procesamiento, la facilidad de conexión y el bajo consumo de energía, así como los anchos de banda limitados, gracias al empleo de señales de voz.



(a) Esquema del sistema IoT (b) Nodo Receptor (c) Nodo Control

Figura 1 – Esquema del sistema IoT y detalle de nodos.

5.2. Funcionamiento del sistema IoT

La Figura 1a muestra un diagrama del funcionamiento del sistema IoT. La comunicación se basa en el protocolo Message Queue Telemetry Transport (MQTT), que transmite información mediante mensajes entre los nodos y el broker MQTT empleando publicación de mensajes y suscripción a temas. Los valores calculados en los nodos se reciben como mensajes publicados en temas definidos en el broker. Los nodos pueden estar suscritos a la información presente en otros temas para recibir información. En nuestro caso, el broker MQTT se implementa en el nodo de control utilizando Eclipse Mosquitto. El protocolo MQTT define 3 niveles de calidad de servicio, de las que hemos seleccionado el más alto, QoS-2, que garantiza la entrega de los mensajes una sola vez, sin pérdidas ni duplicidades.
En cuanto a la seguridad, también hemos implementado el paquete más completo utilizando nombre de usuario y contraseña, tanto en el broker como en los clientes, y un cifrado basado en la certificación SSL para los datos transmitidos. Como los nodos pueden crear un nuevo tema conociendo la dirección del broker y teniendo los permisos necesarios, con sólo publicar en él, para añadir más nodos al sistema IoT sólo tienen que publicar en el tema creado, lo que facilita enormemente la escalabilidad del sistema.
Los datos recibidos de los parámetros calculados van acompañados de una marca temporal y se almacenan localmente en una base de datos (implementada con Mongo DB), a la que podemos acceder para visualizar los datos. Estos datos también pueden ser almacenados en la nube, proporcionando una copia de seguridad y una seguridad extra contra la pérdida de datos.

Para visualizar gráficamente los datos, se deben indicar las posiciones de los nodos y las dimensiones de la sala a analizar, lo que permite además realizar una visualización en forma de mapas de calor, como se observa en la Figura 3, con la predicción del parámetro SII en una de las aulas analizadas.

6. EVALUACIÓN Y RESULTADOS

En esta sección describiremos los resultados obtenidos al evaluar el error en predicción de los parámetros acústicos del conjunto seleccionado (RT60, C50, C80, STI y SII), mediante el modelo de CNN entrenado. Se evalúa el rendimiento del modelo CNN sobre la partición de prueba independiente. Para completar nuestra evaluación, analizamos el tiempo de respuesta utilizando diferentes plataformas de hardware, incluyendo dispositivos típicos empleados en redes IoT. Una vez embebido el modelo de CNN en nuestro sistema IoT, analizamos los resultados obtenidos en una prueba de campo analizando una sala real.

6.1. Precisión en predicción

Los resultados presentados en esta sección analizarán el rendimiento del modelo CNN con respecto a su capacidad para estimar los parámetros de referencia subyacentes a partir de una señal de voz de entrada dada. Como ya se ha mencionado anteriormente, al final del proceso de entrenamiento, la red alcanzó un MSE de 0,08572 sobre el conjunto de datos de validación. Sin embargo, para evaluar el rendimiento sobre un conjunto de datos completamente independiente, hemos testeado el modelo considerando la partición de prueba, que incluye como se ha descrito señales de habla y las respuestas que no han participado de forma alguna en el proceso de entrenamiento. Al hacerlo, obtenemos una visión más precisa de la precisión real de la CNN.
Los resultados de dicha evaluación se recogen en la Tabla 2, donde se muestran el MSE, el Error Medio Absoluto (MAE), el Error Medio Relativo (MRE) y el coeficiente de correlación de Pearson (ρ) para cada parámetro. En este caso, como era de esperar, al utilizar audios de un conjunto de datos independiente, el MSE general aumenta ligeramente hasta 0,0896, en las predicciones de estos 5 parámetros, con un valor medio del coeficiente de correlación superior a 0,99. Luego, el rendimiento final sobre el conjunto de pruebas no se desvía significativamente del obtenido sobre los datos de validación y el modelo parece comportarse correctamente sobre los datos desconocidos. En términos de MRE, el error para todos los parámetros está siempre por debajo del 10 %, con el peor caso dado en C80 (9,29%), y cerca del 2% para STI y SII, por lo que la precisión alcanzada puede ser suficiente para las aplicaciones comunes de monitorización acústica, ya que la desviación estándar de los errores (Error σ) tiende a estar por debajo de la mínima diferencia perceptible para estos parámetros.

Tabla 2: Evaluación de error de predicción de la CNN sobre el conjunto de datos de Test

Parámetro (unidad)	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 (-)	0.0003	0.0132	2.19	0.0135	0.9919
SII (-)	0.0008	0.0205	2.28	0.0211	0.9697
Media	0.0896	0.1339	5.43	0.1732	0.9903

6.2. Rendimiento en tiempo de cálculo

En el caso de los parámetros acústicos de sala, puede parecer que no es imprescindible poder trabajar en tiempo real, debido a que las características acústicas de una sala no varían sustancialmente si no cambia su geometría o los materiales que la componen. Sin embargo, un cambio en el aforo presente puede afectar a diferentes parámetros y por ejemplo en el caso de la inteligibilidad, para evaluarla en diferentes posiciones empleando pocos sensores, o de forma continua, es muy importante poder obtener estos parámetros lo más rápido posible. La velocidad

de cálculo es esencial para los nodos de una red IoT al disponer de capacidad de procesado limitada. Para realizar esta prueba, se han tomado las 1000 señales de audio del conjunto de prueba y medido el tiempo necesario para calcular los parámetros directamente. A continuación, se ha evaluado el tiempo empleado por la CNN en predecir los mismos parámetros. 7 Para obtener datso sobre diferetes dispositivos se ha repetido esta prueba en 4 plataformas diferentes: 2 ordenadores personales y 2 SBC con las siguientes especificaciones técnicas:

- PC-1 (Sobremesa): Intel Core(TM) i7-7700 CPU @ 3.60 GHz, 32 GB RAM, 1 TB HDD.
- PC Portátil: Intel Core(TM) i7-1065G7 CPU @ 1.3 GHz, 16 GB RAM, 500 GB SSD.
- Udoo X86 II-Ultra: Intel 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 c10.

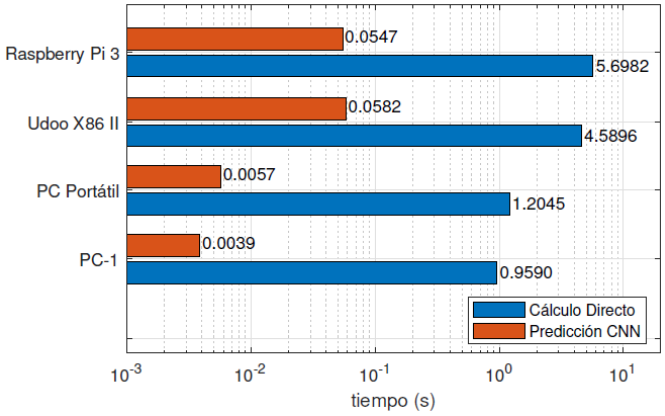


Figura 2: Tiempo empleado en cálculo directo vs. predicción sobre diferentes plataformas

En la Figura 2 podemos ver que, empleando el un ordenador de sobremesa, el modelo CNN permite predecir los parámetros acústicos de la sala en un tiempo medio de 0,0039 segundos, frente a los 0,9590 segundos de tiempo empleado por el cálculo directo. Esto demuestra que la predicción mediante nuestro modelo es 245,8 veces más rápida que el cálculo directo. En el ordenador portátil, la relación desciende un poco hasta 211,3 veces, pero sigue habiendo una ventaja considerable. En el caso de los dispositivos SBC, debido a las menores especificaciones obtenemos tiempos más elevados tanto en el cálculo directo como en la predicción, pero aun así se demuestra una clara ventaja del enfoque basado en CNN, ya que es 78,8 veces más rápido en una Udoo y hasta 104,1 veces más rápido en la Raspberry Pi. Este aumento en la velocidad representa una clara ventaja sobre el cálculo directo, teniendo en cuenta que la obtención de la RI para el cálculo de ciertos parámetros es más compleja y requiere más tiempo, habilitándose además la monitorización continua.

6.3. Prueba de campo en un aula real

Esta prueba tiene como objetivo evaluar el rendimiento del sistema completo AI-IoT sobre aulas reales de las instalaciones de la Escuela Técnica Superior de Ingeniería de la Universitat de València (ETSEUV). El sistema probado se compone de 6 nodos receptores y 1 nodo de control, colocados tal y como se muestra en la Figura 3 para el caso del aula 1.1.3. Para calcular los valores reales de referencia de los diferentes parámetros, se realizaron previamente mediciones de los mismos en las mismas posiciones.
Para esta prueba se han empleado señales de habla no utilizadas hasta ahora, repitiéndose el proceso 10 veces para promediar varias predicciones y obtener una medición más fiable. La Tabla 3 muestra los resultados obtenidos en las predicciones frente a los valores calculados por posición junto con el error medio absoluto (MAE) para las 6 posiciones. Como era de esperar, el error de predicción es algo mayor que el obtenido al evaluar la partición de prueba, debido probablemente a que las aulas son salas nuevas desconocidas, al ruido propio del micrófono, ruido ambiental de fondo u otros sonidos superpuestos. En cualquier caso, las predicciones

obtenidas en esta prueba fueron precisas y eficientes para nuestra aplicación, como se puede ver en la mencionada tabla.
En la Figura 3 se puede apreciar de una manera clara e intuitiva, sobre el mapa de calor de SII cómo cambia la inteligibilidad en función de la distancia al orador, gracias a la monitorización de diferentes puntos en un aula. Esto confirma la validez del marco propuesto para la monitorización mediante IA-IoT de los parámetros acústicos y la inteligibilidad de una sala, permitiendo obtener una descripción significativa del comportamiento acústico de una sala de forma sencilla.

Tabla 3: Valores calculados y predicciones en prueba de campo del sistema AI-IoT.

	RT60 (s)		C50 (dB)		C80 (dB)		STI		SII	
Nodo	Calc.	Pred.	Calc.	Pred.	Calc.	Pred.	Calc.	Pred.	Calc.	Pred.
1	0.85	0.84	3.56	2.81	6.53	5.63	0.64	0.64	0.74	0.66
2	0.84	0.80	3.04	2.56	6.18	5.39	0.65	0.64	0.73	0.64
3	0.77	0.91	3.52	3.28	7.06	6.25	0.65	0.65	0.73	0.71
4	0.81	0.86	3.27	3.64	6.85	6.60	0.66	0.66	0.75	0.72
5	0.81	0.82	3.28	3.27	6.15	6.15	0.66	0.65	0.76	0.71
6	0.79	0.83	4.30	4.24	7.44	7.28	0.67	0.67	0.77	0.77
MAE	0.0483		0.3183		0.4850		0.0033		0.0450	

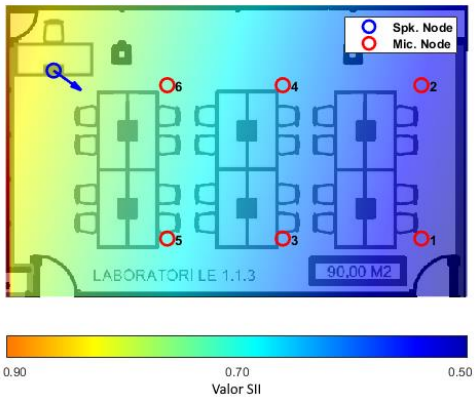


Figura 3: Mapa de calor para SII en prueba de campo del sistema AI-IoT.

7. CONCLUSIONES

En este trabajo, se ha introducido y desarrollado marco completo de IA-IoT para estimación ciega de un conjunto de parámetros acústicos y de inteligibilidad de la sala (RT60, C50, C80, STI y SII) directamente a partir de señales de voz. El sistema propuesto se basa en un modelo de red neuronal convolucional (CNN) entrenado con un conjunto de señales de voz creado mediante salas sintéticas y reales. El modelo resultante consigue un error relativo medio inferior al 5,5% sobre los datos de prueba demostrando ser hasta 245 veces más rápido que el cálculo directo. Además, el modelo se ha integrado en un sistema AI-IoT que permite la monitorización continua y simultánea de los parámetros acústicos en diferentes ubicaciones sin despliegues complejos. Así, el sistema completo ofrece una solución precisa, flexible y eficiente desde el punto de vista computacional para aplicaciones de monitorización acústica a un coste reducido.

AGRADECIMIENTOS

Los autores agradecen a la Agencia Estatal de Investigación (AEI) y al Fondo Europeo de Desarrollo Regional (FEDER) la financiación parcial de esta investigación dentro de los proyectos con referencias PID2021-126823OB-I00 y RTI2018-097045-B-C21, financiados por MCIN/AEI/10.13039/501100011033, por la "Unión Europea NextGenerationEU/PRTR", y por

"FEDER Una forma de hacer Europa" y la subvención BES-2017-082340 financiada por MCIN/AEI/ 10.13039/501100011033 y por "FSE Invirtiendo en tu futuro". También a la Universitat de València, por la ayuda a acciones especiales UV-INV-AE-1544281 y la ayuda para estancias UV-INV_EPDI-1993235.

REFERENCIAS

[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? 43:41–48, 2015.

[2] ISO. ISO 3382. Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces; Part 2: Reverberation time in ordinary rooms; Part 3: Open plan offices. Technical report, UNE-EN, March 2016.

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

[4] 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, June 1997.

[5] Yoichi Ando. Concert Hall Acoustics, volume 17. Springer-Verlag, Berlin, Heidelberg, 1 edition, 7 1985.

[6] 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.

[7] M. F. Duarte and Y. Hen Hu. Vehicle classification in distributed sensor networks. J. Parallel Distrib.Comput., 64:826838, 2004.

[8] Wangyang Yu and W. Bastiaan Kleijn. Room acoustical parameter estimation from room impulse responses using deep neural networks. IEEE/ACM Trans. Audio, Speech and Lang. Proc., 29:436–447, jan 2021.

[9] 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.

[10] Constantinos Papayiannis, Christine Evers, and Patrick A. Naylor. End-to-end classification of reverberant rooms using dnns. IEEE/ACM Trans. Audio, Speech and Lang. Proc., 28:3010–3017, jan 2020.

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

[12] 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 Trans. Audio, Speech and Lang. Proc., 24(4):719–732, apr 2016.

[13] Choi Ling Coriolanus Lam. Improving the speech intelligibility in classrooms. PhD thesis, Dept. of Mechanical Engineering, The Hong Kong Polytechnic University, The address of the publisher, 2010.

[14] R.R McNeer, C.L. Bennett, D.B. Horn, and Dudaryk R. Factors affecting acoustics and speech intelligibility in the operating room: Size matters. Anesth Analg., 124(6):1978–1985.

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

[16] 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.