TECNIACUSTICA ELCHE · 2022

53º CONGRESO ESPAÑOL DE ACÚSTICA
XII CONGRESO IBÉRICO DE ACÚSTICA

APLICACIÓN DE TÉCNICAS DE EXPLICABILIDAD EN CONJUNTOS DE DATOS HTRF UTILIZANDO REDES NEURONALES CONVOLUCIONALES

PACS: 43.00.00 ACOUSTICS, 43.60.−C ACOUSTIC SIGNAL PROCESSING, 07.05.MH NEURAL NETWORKS, FUZZY LOGIC, ARTIFICIAL INTELLIGENCE

Juan Antonio DE RUS[(1)], Aaron LOPEZ-GARCIA[(1)], Ana M. TORRES-ARANDA[(2)], Francesc J. FERRI[(1)], Mario MONTAGUD[(1)] and Maximo COBOS[(1)]

(1)Dpto. Informatica, Universitat de Valencia, Spain, juan.rus@uv.es

(2)Dpto. IEEAC, Universidad de Castilla-La Mancha, Spain

Palabras Clave: Audio espacial, HRTF, XAI, aprendizaje profundo, redes neuronales convolucionales

ABSTRACT.

Spatial audio based on binaural reproduction is one of the major trends in audio technology. To provide a listener with a realistic sensation over headphones, signals are typically filtered with Head-Related Transfer Functions (HRTFs). These describe the transfer properties of sound waves as they travel from the source to the ear canal in free space. Since HRTFs are highly individual (they depend on a subject's anthropometric features), deviations from the user's own HRIRs can affect negatively the listening experience. Therefore, the correct identification of relevant localization cues and their preservation is a topic of continuous interest. In this context, while numerous studies have been carried out in the past to identify salient localization cues, some recent works are exploiting the feature learning capabilities of deep learning-based approaches. In this work, we explore the use of common explainable artificial intelligence (XAI) techniques on convolutional neural networks (CNN), such as Class Activation Mapping (CAM) trained to classify HRTFs into different directional sectors.

RESUMEN.

La reproducción binaural es hoy una técnica popular para el audio inmersivo. Para una sensación realista a través de auriculares, las señales sonoras suelen filtrarse utilizando las funciones de transferencia relacionadas con la cabeza (HRTFs). Éstas describen las propiedades de transferencia de las ondas sonoras cuando se desplazan desde la ubicación de una fuente sonora determinada hasta el canal auditivo en el espacio libre. Dado que las HRTF tienen un carácter individual (dependen de las características antropométricas del sujeto), las desviaciones sobre las propias HRTF de un usuario particular pueden afectar negativamente a su experiencia. Por tanto, la correcta identificación de las componentes relevantes para la localización y su preservación en la síntesis es un tema de continuo interés. En este contexto, mientras que en el pasado se han llevado a cabo numerosos estudios para identificar qué componentes frecuenciales son significativas en la localización, algunos trabajos recientes están explotando las capacidades de aprendizaje de los enfoques basados en redes neuronales profundas. En este trabajo, exploramos el uso de técnicas comunes de inteligencia artificial explicable (XAI) en redes convolucionales, como el mapeo de activación de clases (CAM), en modelos entrenados para clasificar conjuntos de datos HRTF en diferentes sectores direccionales.

1. INTRODUCCIÓN

Con la llegada de los escenarios acústicos inmersivos para la realidad virtual, lograr una localización precisa de las fuentes de sonido se ha convertido en un gran desafío. Una de las peculiaridades que complica el desarrollo de sistemas de reproducción precisos es que la audición espacial humana está íntimamente relacionada con la anatomía del oyente. Los efectos acústicos causados por la cabeza, el torso, los hombros y el pabellón auricular tienen un gran impacto en la capacidad de localización humana. Es un hecho bien conocido que tales características humanas pueden ser descritas estadísticamente para ayudar al proceso de localización auditiva [1]. La mayoría de los modelos acústicos suelen utilizar Respuestas de Impulso Relacionadas con la Cabeza (HRIR) en el dominio del tiempo o, de forma equivalente, Funciones de Transferencia Relacionadas con la Cabeza (HRTF) en el dominio de la frecuencia. Con respecto a las señales HRTF, varios trabajos han propuesto el uso de técnicas de preprocesamiento y posprocesamiento para capturar la influencia relativa de características antropométricas relevantes [2]. La localización de fuentes virtuales también puede mejorarse mediante el escalado de las funciones de transferencia direccional [3]. Un hallazgo interesante que coincide con el objetivo de este trabajo es que las prominencias (ya sean picos o muescas) en las curvas de las funciones de transferencia pueden revelar información asociada a la elevación de una fuente [4] [5].

El uso de redes neuronales profundas para la ayuda a la toma de decisiones en entornos acústicos ha ido en constante crecimiento. En particular, el uso de redes neuronales convolucionales (CNN) es crucial en la mayoría de las tareas de soporte artificial, como la clasificación de escenas acústicas [6], el etiquetado de música [7] y el reconocimiento de emociones del habla [8], entre muchas otras. En la misma línea, el desarrollo de conjuntos de datos espaciales binaurales permite entrenar redes neuronales para modelar HRTF personalizados y permitir una experiencia auditiva más realista. Estudios recientes han demostrado la posibilidad de capturar características de audio espacial en conjuntos de datos HRTF usando [9] de CNN. Estos estudios pueden entenderse como una nueva forma de estudiar el desempeño humano en la localización de fuentes. Otros estudios se han centrado recientemente en la discriminación anverso-reverso de las grabaciones de música binaural [10], lo que ha revelado información interesante sobre las bandas de frecuencia relevantes que contribuyen a dicha tarea de discriminación.

En el mismo espíritu de [9], este artículo se centra en la aplicación de técnicas de inteligencia artificial explicable (XAI) para el análisis de conjuntos de datos HRTF. Más específicamente, consideramos el análisis del conjunto de datos HRTF CIPIC [11] usando un modelo 1D-CNN convencional. A diferencia de [9], los HRTF transformados a escala de mel se utilizan directamente como entrada a la red. Además, aplicamos dos técnicas diferentes de XAI para el análisis de prominencia, a saber, el mapeo de activación de clases (CAM) [12] y el método CAM más general basado en gradientes (Grad-CAM) [13]. Los resultados obtenidos con estas técnicas se analizarán para descubrir bandas de frecuencia en las HRTF que codifican claves de ubicación relevantes para determinar la elevación de una fuente.

2. CASO DE ESTUDIO

2.1 Conjunto de datos experimentales y preprocesamiento de datos

La base de datos CIPIC HRTF es una base de datos de dominio público de mediciones HRTF de alta resolución espacial para 45 sujetos diferentes, incluido el maniquí KEMAR con pinnas pequeñas y grandes [11]. Incluye 2.500 mediciones de HRIR para 45 sujetos en 25 acimuts diferentes y 50 elevaciones diferentes (1250 direcciones) en incrementos angulares de aproximadamente 5º. Las medidas estándar se registraron en 25 acimuts interaurales polares diferentes y 50 elevaciones interaurales polares diferentes. La duración de la muestra de cada HRIR es de 200 muestras (4,5ms a una frecuencia de muestreo de 44,1 kHz). Se realizaron medidas especiales adicionales del maniquí KEMAR para los planos frontal y horizontal.

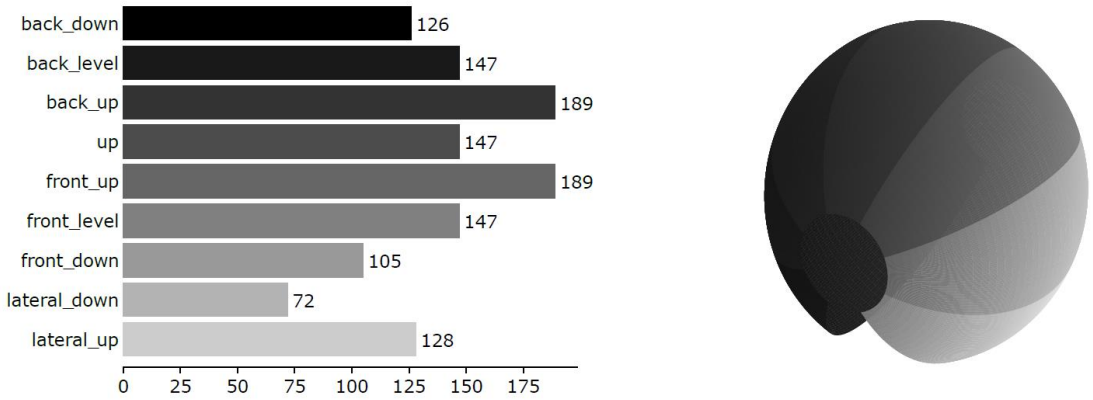

Figura 1 - Ubicación de la fuente espacial en el conjunto de datos CIPIC dividida en 9 regiones para la tarea de clasificación.

Al igual que en [9], enfocamos este estudio en el análisis de señales de elevación. Con este fin, dividimos todas las respuestas de la base de datos CIPIC en un conjunto de 9 regiones esféricas según su elevación, como se muestra en la **¡Error! No se encuentra el origen de la referencia.**. Nótese que el número de direcciones muestreadas dentro de cada una de las clases de elevación no es uniforme, aunque no está severamente desequilibrado.

El conjunto de datos completo se compone de 56250 muestras HRIR (correspondientes a la combinación de los 45 sujetos, 25 ángulos de acimut y 50 ángulos de elevación) con sus canales ipsilateral y contralateral asociados.

Para obtener las señales HRTF correspondientes, calculamos la transformada rápida de Fourier unilateral de cada canal con 512 puntos, lo que da como resultado 257 intervalos de frecuencia. Para proporcionar a la red una entrada motivada por la percepción, deformamos el eje de frecuencia considerando un mapeo a escala de mel. Esto se logra dividiendo el rango [0-22050] Hz en 257 puntos espaciados uniformemente en la escala de mel, y tomando los intervalos de frecuencia que están más cerca de sus frecuencias equivalentes en Hz. Finalmente, solo se considera el espectro de magnitud de cada canal en escala logarítmica. Por lo tanto, la forma de cada ejemplo de entrada es 257×2 .

El conjunto de datos se dividió en dos particiones para entrenamiento y validación. Los datos de 36 sujetos (45000 muestras) se utilizaron para la partición de entrenamiento y los datos de 9 sujetos diferentes (11250 muestras) para la validación.

2.2 Arquitectura del modelo

Este trabajo considera un modelo completamente convolucional con una arquitectura sencilla, representada en la Figura 2. El diseño de la red se llevó a cabo con el objetivo de lograr una clasificación de precisión significativa manteniendo el modelo lo suficientemente simple para facilitar la aplicación de técnicas XAI comunes. Consiste en tres bloques convolucionales 1D con activación ReLU y max-pooling entre bloques para reducir la frecuencia de la información. Se utiliza una última capa convolucional seguida de Global Average Pooling (GAP) para resumir las respuestas del filtro antes de la capa densa final, configurada con activación softmax.

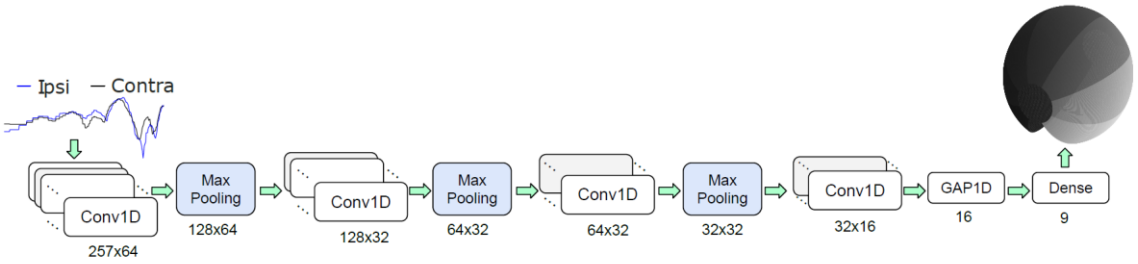

Figura 2 - Topología de la arquitectura convolucional desarrollada para este artículo para clasificar los HRTF en nueve sectores de elevación.

La información relacionada con la construcción de cada capa de la CNN se muestra en la Tabla 1. Muestra información detallada sobre cómo se construyó el modelo, incluidas las dimensiones involucradas a diferentes profundidades y la cantidad de parámetros asociados.

Tabla 1 - Configuración de las capas CNN.

	Input	Conv1	MPool1	Conv2	MPool2	Conv3	MPool3	Conv4	GAP	Dense
Dimensión	257	257	128	128	64	64	32	32	16	9
Filtros	2	64	64	32	32	32	32	16	1	1
Tamaños	0	64	2	16	2	8	2	8	0	0
Parámetros	0	2112	0	32800	0	8224	0	4112	0	153

El modelo fue entrenado con el optimizador de $Adam(\eta = 10^{-3})$ usando entropía cruzada categórica como función de pérdida. Para la adaptación establecimos un lote de 16 elementos y 100 épocas con parada anticipada, que se activó durante los procesos de entrenamiento.

2.3 Explicabilidad de características

El uso de técnicas XAI puede entenderse como una fase adicional en la evaluación del modelo [14]. Como una forma de dar explicabilidad de la misma manera que los estudios mencionados anteriormente, hemos analizado los resultados de nuestra arquitectura convolucional profunda. Una vez que nuestro modelo esté ajustado y tenga un rendimiento aceptable, queremos eliminar la caja negra producida por la arquitectura CNN. Con ese propósito, hemos aplicado CAM [12] y Grad-CAM [13] para analizar mapas de saliencia y mapas de activación de clases. Aunque ambas técnicas XAI se diseñaron únicamente para cubrir las limitaciones de las CNN, su implementación está ampliamente extendida gracias a su capacidad para generar mapas de localización que destacan regiones espaciales significativas [15]. CAM surgió originalmente como una técnica que produce mapas de activación de clases de CNN aplicados a tareas de detección de objetos. Permitió que los modelos entrenados localizaran patrones específicos de clase en la detección de imágenes. Debido a que la principal desventaja de CAM es el requisito de una capa GAP para operar sobre los filtros convolucionales, Grad-CAM supera esa limitación mediante el uso de gradientes sobre la primera capa densa de la arquitectura. Independientemente de sus diferencias, ambas técnicas se basan en la suposición de que el vector de decisión Y^c para la clase c se describe mediante los mapas de características A^k de la última capa convolucional. Entonces, el mapa de saliencia para la clase c es una agregación ponderada de los componentes espaciales de cada mapa de características, escrito como:

$$\text{Class} - \text{score: } Y^c = \sum_k w_k^c \sum_i \sum_j A_{ij}^k, \qquad \text{Class} - \text{Saliency: } L_{ij}^c = \sum_k w_k^c A_{ij}^k \tag{1}$$

Donde los componentes i y j denotan los índices matriciales de los componentes espaciales y el índice k asocia los filtros numéricos. Con respecto al esquema de ponderación (w_k^c), CAM utiliza una proyección hacia atrás de los pesos de salida en los mapas de la última capa convolucional. Por el contrario, Grad-CAM estima cada peso de clase para cada mapa de características como una combinación lineal de las derivadas parciales $\frac{\partial Y^c}{\partial A_{ij}^k}$. Su implementación está muy relacionada con la clasificación de imágenes [12] [13] [15] que van desde pruebas clínicas [16] hasta simulación de juegos de computadora a través del aprendizaje por refuerzo [17]

3. RESULTADOS

3.1 Rendimiento de la CNN

Después del entrenamiento, el modelo logró una precisión global de 0.8090 en el conjunto de validación, lo que sugiere que el modelo funciona razonablemente bien en la tarea de determinar la clase de elevación de una HRTF determinada, independientemente de su azimut. Para analizar el rendimiento de la clasificación con más detalle, la Figura 3 muestra la matriz de confusión (porcentaje de aciertos) para las diferentes clases de elevación. La proporción de ejemplos de clase se muestra en escala roja. Nótese que la mayoría de las predicciones incorrectas se clasifican erróneamente en regiones espaciales adyacentes, lo que muestra robustez en términos de comprensión de las señales espaciales subyacentes.

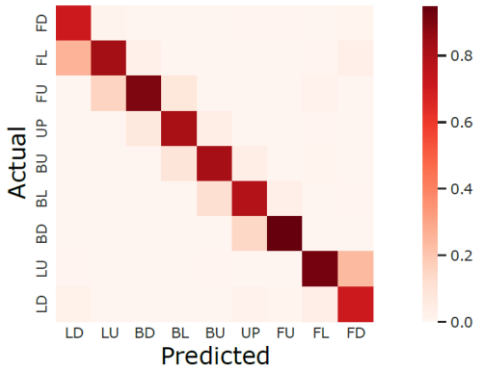

Figura 3 - Matriz de confusión calculada por la proporción de aciertos para el modelo CNN.

3.2 Mapas de saliencia

En esta sección se muestran los resultados obtenidos por las técnicas XAI consideradas en este trabajo, analizando los mapas de prominencia proporcionados por los métodos CAM y Grad-CAM. Estos mapas de saliencia proporcionan información significativa sobre cómo los núcleos convolucionales procesan un ejemplo de entrada dado para determinar su clase de elevación correspondiente. La Figura 4 muestra respuestas HRTF representativas seleccionadas del conjunto de validación junto con su saliencia correspondiente en segundo plano. La selección se hizo según un criterio de máxima probabilidad, es decir, corresponden a las respuestas pertenecientes a cada clase que el modelo clasificó con mayor confianza. Las zonas rojas oscuras corresponden a frecuencias que tienen una alta saliencia, mientras que las zonas claras corresponden a frecuencias que son menos relevantes para la tarea de clasificación. Para cada clase, la saliencia de CAM se muestra en la parte superior, mientras que el resultado de Grad-CAM se muestra en la parte inferior. Puede observarse para cada uno de los ejemplos representados que tanto CAM como Grad-CAM proporcionan zonas de saliencia similares,

aunque la intensidad de dichas zonas puede variar de un método a otro. En general, existe un alto nivel de acuerdo en los resultados proporcionados por ambos enfoques.

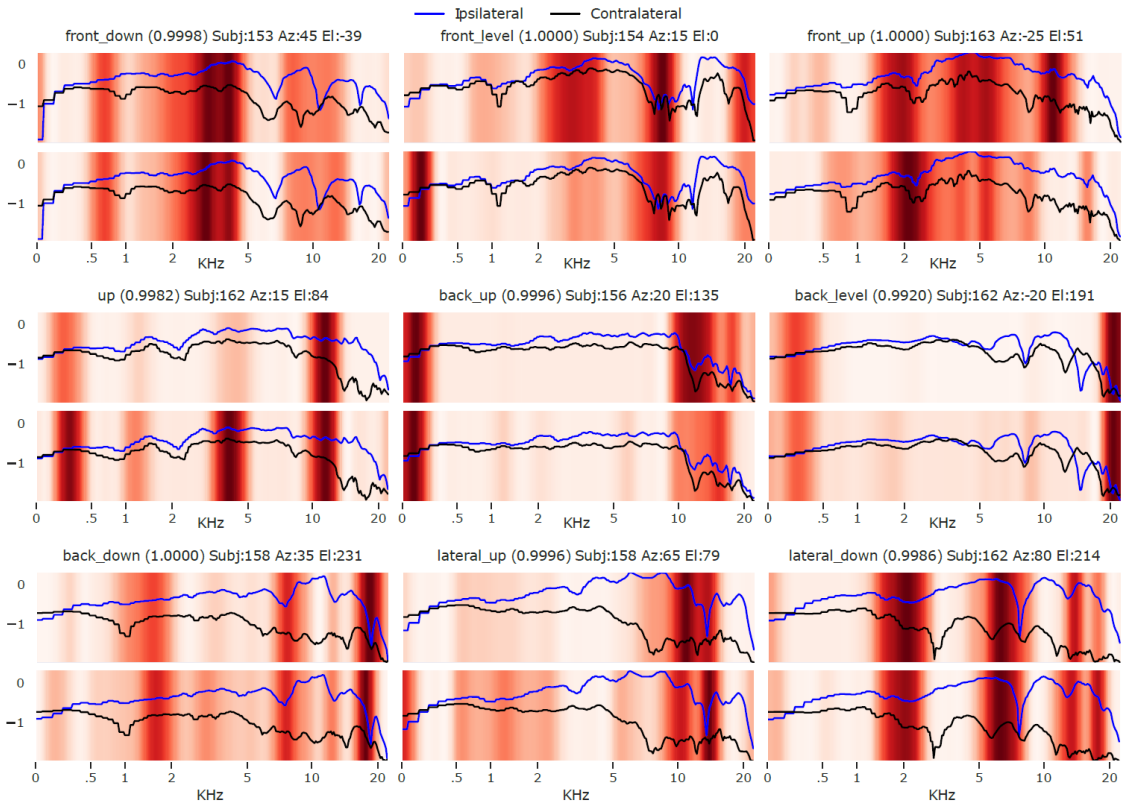

Figura 4 - Mapas de saliencia CAM (arriba) y Grad-CAM (abajo) sobre la muestra más representativa de cada clase. La probabilidad de clase pronosticada $\hat{y}_i$ se da entre paréntesis. La barra de color está en escala roja, luego los tonos blancos indican baja

Para obtener más información sobre los componentes espaciales relevantes procesados por el modelo CNN, analizamos los mapas de saliencia promedio resultantes de CAM y Grad-CAM en todos los sujetos y ángulos de acimut. El objetivo es obtener una imagen general de las frecuencias relevantes que ayuden a la clasificación de todo el conjunto de datos. Las regiones laterales no fueron consideradas ya que realmente no afectan la elevación. Los mapas promedio de la saliencia para elevación-frecuencia se muestran en la Figura 5, donde las zonas iluminadas indican más importancia y las oscuras menos importancia para la clasificación.

Finalmente, Figura 6 muestra como una imagen las prominencias de las respuestas individuales pertenecientes a cada clase en el conjunto de validación, tanto para CAM como para Grad-CAM. Nótese que existe una alta correlación entre los resultados de saliencia obtenidos para las diferentes respuestas dentro de cada clase, sugerida por bandas verticales prominentes ubicadas en regiones de frecuencia más estrechas o anchas. Nuevamente, los resultados obtenidos por CAM y Grad-CAM son considerablemente similares.

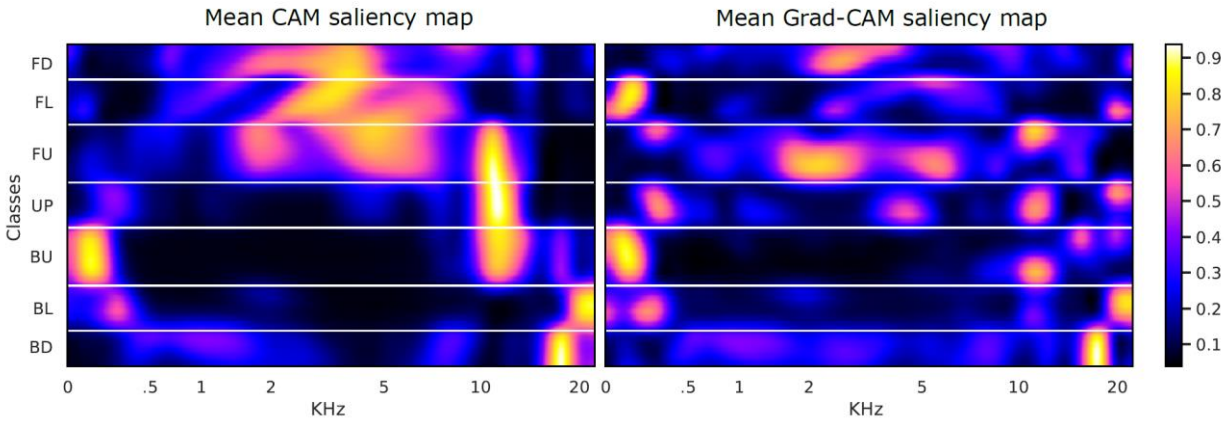

Figura 5 - Mapas de saliencia CAM (izquierda) y Grad-CAM (derecha) promediados entre sujetos y azimut, mostrados por clase de elevación.

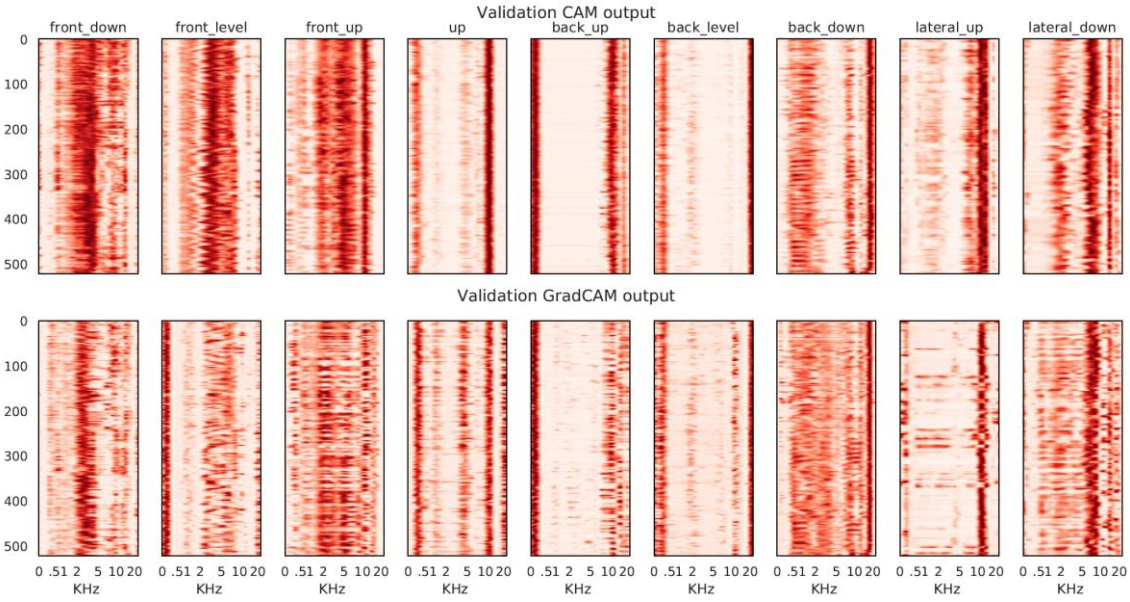

Figura 6 - Bandas de prominencia de CAM (arriba) y Grad-CAM (abajo) por clase en la partición de validación para predicciones correctas. La barra de colores está escalada, los tonos claros indican baja relevancia y los rojos oscuros alta relevancia.

3.3 Discusión

Los mapas de saliencia obtenidos en la sección anterior están considerablemente de acuerdo con algunos de los resultados derivados de experimentos psicoacústicos previos. A continuación, describimos cómo nuestros resultados están vinculados a algunas señales espectrales conocidas y efectos relacionados con la elevación y la discriminación frontal-posterior. Tenga en cuenta que, dado que el mapa de saliencia promedio de CAM en la Figura 5 (izquierda) parece ser más consistente que el obtenido de Grad-CAM, discutiremos nuestros hallazgos teniendo en cuenta principalmente los resultados de CAM.

Muchos estudios han demostrado que las distorsiones espectrales causadas por pinnas en el rango de alta frecuencia aproximadamente por encima de 4 kHz actúan como señales para la localización en el plano medial [18]. De hecho, al observar los mapas de saliencia resultantes, las saliencias más intensas se encuentran, como se esperaba, en el rango de frecuencia media y alta. Sin embargo, hay algunos efectos interesantes de baja frecuencia (por debajo de 500 Hz) en las clases “up”, “back-up”, “back-level” y “back-down”. Estos pueden sugerir que los HRTF de las direcciones posteriores muestran algunas características de baja frecuencia que ayudan a la percepción de la elevación que pueden no estar presentes en las direcciones frontales. En este contexto, aunque [4] informó que una señal de posición posterior (back) aparece como un pequeño pico alrededor de 12 kHz, observamos que la saliencia promedio es alta en esta frecuencia para “back up”, pero también las contribuciones de frecuencias más altas en “back level” y “back down”, lo que sugiere un cambio de señales de localización hacia frecuencias muy altas cuando una fuente posterior se mueve de abajo hacia arriba.

Butler y Belendiuk [19] demostraron que la muesca prominente se mueve hacia las frecuencias más bajas a medida que la fuente de sonido se mueve desde arriba hacia abajo del eje auditivo en la mitad frontal del plano medial. Para las direcciones frontales, moviéndose de “front-up” a “front-down”, podemos ver el cambio de saliencia hacia frecuencias más bajas (desde aproximadamente 8 kHz hacia los 3 kHz).

Además, como se observa en la Figura 4 y Figura 6, parece haber saliencias complementarias en muestras de clases espaciales opuestas que pueden insinuar más señales, como se muestra a continuación. Las muestras de las clases “lateral-up” y “lateral-down” presentan saliencias opuestas en la banda de 8-15 kHz, mientras que son similares en el resto de frecuencias. De manera similar, las muestras de las clases “front-level” y “back-level” presentan saliencias opuestas en la banda de 2-10 kHz, mientras que son bastante similares fuera de este rango.

4. CONCLUSIONES

Este trabajo presentó un estudio preliminar sobre el uso de técnicas de inteligencia artificial explicables, a saber, CAM y Grad-CAM, para ayudar en la interpretación de las señales de elevación de HRTF. Con este fin, entrenamos una red neuronal convolucional en respuestas HRTF deformadas a escala de mel extraídas de la base de datos CIPIC. El modelo fue entrenado para clasificar las respuestas en 9 clases espaciales diferentes relacionadas con diferentes sectores de elevación, y mostró capacidades de generalización considerables sobre un conjunto de validación con respuestas de sujetos diferentes a los del conjunto de entrenamiento. Las técnicas de explicabilidad se aplicaron sobre el modelo entrenado para obtener mapas de saliencia que indican las bandas de frecuencia relevantes utilizadas por la red para clasificar una muestra de entrada dada en una de las clases de elevación. Aunque tanto CAM como Grad-CAM proporcionaron regiones de saliencia similares, los resultados de CAM parecieron ser más consistentes en los conjuntos de entrenamiento y validación. Las regiones de saliencia identificadas por las técnicas explicables aplicadas también fueron consistentes con la mayoría de los hallazgos obtenidos a través de experimentos psicoacústicos, aunque se obtuvieron efectos de baja frecuencia inesperados adicionales para direcciones traseras.

AGRADECIMIENTOS

Este trabajo recibió financiación de Grant RTI2018-097045-B-C21 financiado por MCIN/AEI/10.13039/501100011033 y ``ERDF A way of making Europe". El trabajo de J.A. De Rus ha sido por financiado por MCIN/AEI/10.13039/ 501100011033, en el marco de la subvención FPU20/05384, y " European Social Fund (ESF) Investing in your future". El trabajo de M. Montagud ha sido por financiado por MCIN/AEI/10.13039/ 501100011033, en el marco de la subvención RYC2020-030679-I, y la ‘‘European Social Fund (ESF) Investing in Your Future.’

REFERENCIAS

[1] C. L. Searle, L. D. Braida, M. F. Davis y H. S. Colburn, «Model for auditory localization,» *The Journal of the Acoustical Society of America,* vol. 60, pp. 1164-1175, 1976.

[2] M. Zhu, M. Shahnawaz, S. Tubaro y A. Sarti, «HRTF personalization based on weighted sparse representation of anthropometric features,» de *2017 International Conference on 3D Immersion (IC3D)*, 2017.

[3] J. C. Middlebrooks, «Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency,» *The Journal of the Acoustical Society of America,* vol. 106, pp. 1493-1510, 1999.

[4] J. Hebrank y D. Wright, «Spectral cues used in the localization of sound sources on the median plane,» *The Journal of the Acoustical Society of America,* vol. 56, pp. 1829-1834, 1974.

[5] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization, The MIT Press, 1996.

[6] J. Abeßer, «A Review of Deep Learning Based Methods for Acoustic Scene Classification,» *Applied Sciences,* vol. 10, 2020.

[7] T. Kim, J. Lee y J. Nam, «Sample-level CNN architectures for music auto-tagging using raw waveforms,» de *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018.

[8] Mustaqeem y S. Kwon, «A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition,» *Sensors,* vol. 20, 2020.

[9] E. Thuillier, H. Gamper y I. J. Tashev, «Spatial Audio Feature Discovery with Convolutional Neural Networks,» de *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[10] S. K. Zieliński, P. Antoniuk, H. Lee y D. Johnson, «Automatic Discrimination between Front and Back Ensemble Locations in HRTF-Convolved Binaural Recordings of Music,» *EURASIP J. Audio Speech Music Process.,* vol. 2022, 2022.

[11] V. R. Algazi, R. O. Duda, D. M. Thompson y C. Avendano, «The CIPIC HRTF database,» de *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, 2001.

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva y A. Torralba, *Learning Deep Features for Discriminative Localization,* arXiv, 2015.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh y D. Batra, «Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,» de *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[14] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins y others, «Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,» *Information fusion,* vol. 58, p. 82–115, 2020.

[15] K. Li, Z. Wu, K.-C. Peng, J. Ernst y Y. Fu, «Tell me where to look: Guided attention inference network,» de *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[16] J. Chang, J. Lee, A. Ha, Y. S. Han, E. Bak, S. Choi, J. M. Yun, U. Kang, I. H. Shin, J. Y. Shin, T. Ko, Y. S. Bae, B.-L. Oh, K. H. Park y S. M. Park, «Explaining the Rationale of Deep Learning Glaucoma Decisions with Adversarial Examples,» *Ophthalmology,* vol. 128, pp. 78-88, 2021.

[17] H.-T. Joo y K.-J. Kim, «Visualization of Deep Reinforcement Learning using Grad-CAM: How AI Plays Atari Games?,» de *2019 IEEE Conference on Games (CoG)*, 2019.

[18] K. Iida y Y. Ishii, «Individualization of the head-related transfer functions on the basis of the spectral cues for sound localization,» de *Principles and Applications of Spatial Hearing*, World Scientific, 2011, pp. 159-178.

[19] R. A. Butler y K. Belendiuk, «Spectral cues utilized in the localization of sound in the median sagittal plane,» *The Journal of the Acoustical Society of America,* vol. 61, p. 1264–1269, 1977.