TECNIACÚSTICA
CUENCA 2023

IMPACTO DEL PREPROCESAMIENTO EN LA CLASIFICACIÓN DE LA ELEVACIÓN DE SEÑALES HRTF EN MÚLTIPLES CONJUNTOS DE DATOS

Juan A. De Rus[1*] *Jesus Lopez-Ballester*[1] *Mario Montagud*[1,2]

Francesc J. Ferri[1] *Jose J. Lopez*[3] *Maximo Cobos*[1]

[1] Computer Science Department, Universitat de Valencia, Spain
[2] i2CAT Foundation, Barcelona, Spain
[3] iTEAM, Universitat Politècnica de València, Spain

RESUMEN

La localización de fuentes de sonido en el plano horizontal se basa en diferencias en el nivel y tiempo interaurales. Sin embargo, identificar la elevación dadas las Funciones de Transferencia Relacionadas con la Cabeza (HRTF) sigue siendo un desafío. Los patrones espectrales desempeñan un papel importante en la localización en la dimensión vertical y son altamente individuales debido a las características anatómicas de cada persona. En un estudio previo, se propuso una red neuronal convolucional simple para clasificar las señales HRTF en sectores de elevación y detectar las señales espectrales relacionadas con la elevación. Aunque el modelo obtuvo buenos resultados, solo se entrenó y validó en la base de datos CIPIC. En este trabajo, nos enfocamos en desarrollar un modelo que pueda generalizarse a múltiples conjuntos de datos de HRTF, logrando un buen rendimiento de clasificación en diferentes sujetos y medidas. Dado que cada conjunto de datos se obtiene en condiciones diferentes, el preprocesamiento de los datos puede tener un impacto significativo en el rendimiento del modelo. Se exploran diferentes técnicas de preprocesamiento y se evalúa su impacto en la clasificación, con el fin de seleccionar estrategias de estandarización adecuadas para trabajar con múltiples conjuntos de datos de HRTF.

ABSTRACT

The process of locating sound sources on the horizontal plane primarily relies on perceived differences in both interaural level and time. However, recognizing elevation signals within Head-Related Transfer Functions (HRTFs) remains a complex task. Spectral signals play a crucial role in identifying the vertical position of sound sources and are highly personalized, influenced by the unique anatomical features of individuals, including the shape of their ears, head, and torso. A prior study introduced a straightforward 1D convolutional neural network (CNN) designed to categorize HRTF signals into distinct elevation categories, thereby identifying spectral elevation signals using interpretability techniques. Despite achieving promising results, this model was exclusively trained and validated using the CIPIC database. This research project centers on the creation of a model capable of generalizing across multiple HRTF datasets, ensuring strong classification performance across various individuals and measurement conditions. Given that each dataset is collected under differing conditions (such as the source signal used, distance between transmitters and receivers, spatial resolution, and calibration), the preprocessing of data could substantially influence the overall performance of inter-dataset models. This study explores different preprocessing techniques and assesses their impact on the classification task to select meaningful standardization strategies.

Palabras Clave— HRTF, señales de elevación, preprocesamiento de conjuntos de datos, redes neuronales convolucionales.

1. INTRODUCCIÓN

Las HRTF describen la transmisión del sonido desde un punto en el espacio hasta el canal auditivo humano [1]. Este concepto encuentra aplicación en diversos campos, como el diseño de sonido personalizado para individuos, incluyendo la cancelación de ruido [2] y la creación de entornos de Realidad Virtual (RV) inmersivos [3]. La HRTF de cada persona es única, influenciada por factores como la forma de la cabeza, el torso, los hombros y las orejas, entre otras características [4].

El objetivo principal de nuestro estudio es desarrollar un modelo de clasificación basado en Redes Neuronales Convolucionales (CNN, por sus siglas en inglés) que utilice eficazmente datos de múltiples conjuntos de datos de HRTF para identificar la ubicación en elevación en el plano medio.

[*] ***Autor de contacto****: juan.rus@uv.es*

Pretendemos explorar qué señales de localización son cruciales para determinar la elevación de una respuesta de HRTF dada, y planteamos la hipótesis de que la capacidad del modelo para clasificar el sector de elevación capturará inherentemente estas características, permitiendo la generalización en todos los conjuntos de datos. Para lograr esto, empleamos técnicas de estandarización de datos y comparamos el rendimiento del modelo en los diferentes conjuntos de datos, resaltando el impacto de las diferencias entre los conjuntos de datos en los resultados.

1.1. Desafíos en la localización de fuentes de sonido en elevación

Para localizar una fuente de sonido en el plano horizontal, se pueden emplear diversas señales de localización, como la diferencia de tiempo interaural (ITD), la diferencia de nivel interaural (ILD), señales espectrales (SC) y directividad del plano horizontal (HPD). Sin embargo, estas señales no son igualmente efectivas para la localización en elevación. En el plano vertical, estamos limitados a utilizar señales espectrales producidas por factores antropométricos, incluyendo reflexiones y refracciones desde la oreja y el torso [5]. Por lo tanto, mientras que ITD y ILD suelen ser suficientes para determinar la localización horizontal de una fuente de sonido, las señales espectrales desempeñan un papel crucial en la determinación de la ubicación en elevación.

Investigaciones previas han demostrado que las distorsiones de la oreja (pinna) se manifiestan como señales espectrales para la localización en elevación más allá de la banda de frecuencia de 4 kHz [6], extendiéndose hasta 10 kHz [7]. Además, a 12 kHz, una señal espectral aparece como un pico, indicando que el sonido proviene desde atrás del oyente [8]. Otro efecto a tener en cuenta es que la muesca significativa se desplaza hacia bandas de frecuencia más bajas a medida que el sonido se mueve desde el cenit hacia la mitad frontal inferior del plano medio [9].

1.2. Trabajo previo

En un estudio anterior [10], entrenamos una CNN para clasificar HRTFs en diversos sectores de elevación que iban desde la zona frontal inferior (Front Down) hasta la posterior inferior (Back Down), así como laterales, como se muestra en la Tabla 1, utilizando datos del conjunto de datos CIPIC [11]. Utilizamos coordenadas esféricas con la convención "side", que utiliza un ángulo lateral que varía en el plano horizontal desde [-90, 90] y un ángulo polar que varía en el plano medio desde [-90, 270]. Elegimos las CNN debido a sus capacidades de reconocimiento de patrones [12], que se han utilizado con éxito para capturar características de audio espacial en HRTFs [13], así como en otras tareas orientadas al sonido, como la clasificación de escenas acústicas [14], etiquetado de música [15], reconocimiento de voz [16] o discriminación automática entre ubicaciones frontales y traseras en grabaciones binaurales [17].

Tabla 1 - Rangos Angulares de cada sector de elevación

Clase	Ángulo Polar	Ángulo Lateral
Front Down	[-90, -20]	[-60, 60]
Front Level	(-20, 20]	[-60, 60]
Front Up	(20, 70]	[-60, 60]
Up	(70, 110]	[-60, 60]
Back Up	(110, 160]	[-60, 60]
Back Level	(160, 200]	[-60, 60]
Back Down	(200, 270]	[-60, 60]
Lateral Up	[0,∞)	|lateral| > 60
Lateral Down	(-∞, 0)	|lateral| > 60

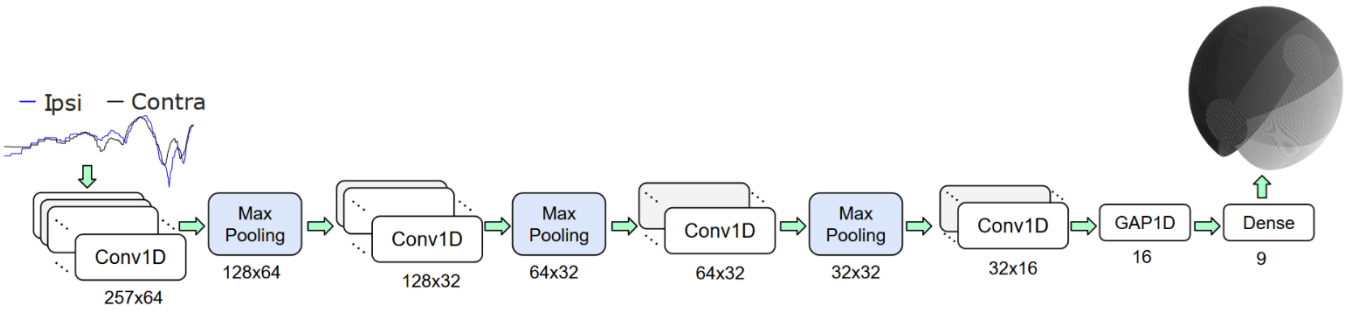


Figura 1 - Topología de la arquitectura de convolución desarrollada en el estudio anterior para clasificar las HRTF en nueve sectores de elevación

Nuestro modelo completamente convolucional consistió en tres bloques convolucionales 1D con activación ReLU y max-pooling entre bloques, seguidos de una última capa convolucional y una capa de pooling promedio global para resumir las respuestas de los filtros antes de la última capa densa con activación softmax (Figura 1). Este modelo simple logró una precisión significativa. Se pueden encontrar más detalles sobre el modelo y su entrenamiento en [10].

Luego, exploramos el uso de técnicas comunes de inteligencia artificial explicables (XAI) [18] para determinar

en qué se fijaba el modelo para hacer sus predicciones, incluyendo qué partes de los datos eran más importantes o relevantes para la predicción. Comparamos los resultados con los de la literatura. Las dos técnicas de XAI que empleamos fueron el Mapeo de Activación de Clase (CAM) [19] y CAM de gradiente (GradCam) [20].

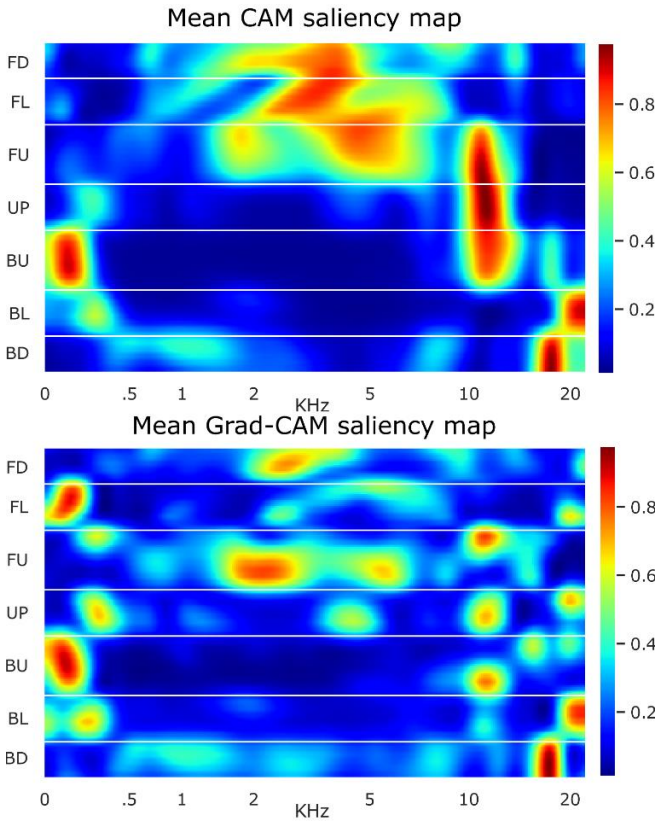


Figura 2 - Mapas de Saliencia CAM (arriba) y Grad-CAM (abajo) promediados entre sujetos y azimut, mostrados por clase de elevación.

1.2.1 Hallazgos

Identificamos bandas de frecuencia significativas que presentaban una mayor prominencia al ser clasificadas en correspondencia con diferentes sectores de elevación en los datos. También encontramos una prominencia alta (lo que podría indicar posibles señales de elevación) en la banda de frecuencia baja por debajo de 500 Hz para las regiones traseras desde "Back-Down" hasta "Up"(hacia arriba), que no se encuentra en las regiones frontales (Figura 2).

Además, encontramos un posible efecto de complementariedad entre regiones opuestas, como "Front Level" y "Back Level", con prominencias opuestas en la banda de frecuencia de 2 a 10 kHz, que son casi idénticas fuera de este rango (Figura 3).

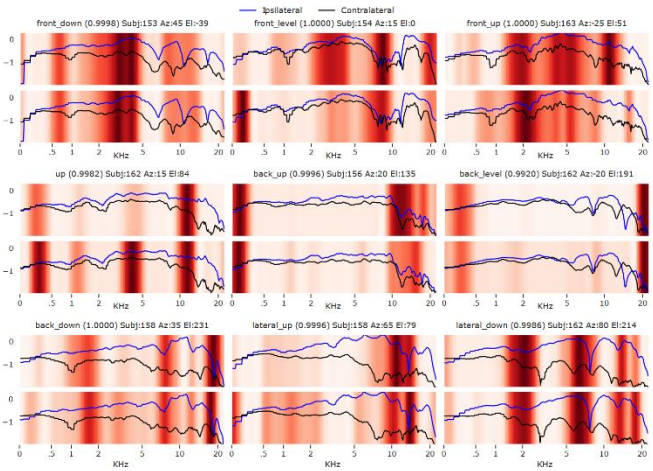


Figura 3 - Mapas de saliencia CAM (arriba) y Grad-CAM (abajo) sobre la muestra más representativa de cada clase. La probabilidad de clase predicha se indica entre paréntesis. La leyenda de colores va desde rojo, indicando alta relevancia, hasta tonos blancos, indicando baja saliencia.

2. PRECISIÓN DEL MODELO ENTRE CONJUNTOS DE DATOS

Como se explicó anteriormente, el enfoque de este trabajo se centra en el desarrollo de modelos para la clasificación de la elevación de HRTFs, con un énfasis particular en los resultados entre conjuntos de datos. En otras palabras, estamos interesados en examinar qué tan bien un modelo funciona cuando se entrena en un conjunto de datos de HRTF pero se prueba en datos de un conjunto de datos diferente.

2.1 Formato de datos utilizado

Para abordar el desafío de trabajar con diferentes conjuntos de datos, cada uno con sus características y particularidades únicas, se han realizado diversos esfuerzos para estandarizar los formatos de datos, como el formato Marl-Nyu para Matlab [21]. En nuestro caso, para facilitar el trabajo con diferentes conjuntos de datos, optamos por utilizar el Formato Espacialmente Orientado para Acústica (SOFA, por sus siglas en inglés) [22]. El Formato SOFA se caracteriza por incluir información autocontenida sobre la descripción de la configuración de medición y todos los elementos relevantes, como el oyente, la fuente y la sala, en cada archivo.

Además, el formato SOFA permite utilizar los datos como HRTFs en el dominio de la frecuencia o como Respuestas de Impulso Relativas a la Cabeza (HRIRs) en el dominio del tiempo. En este trabajo, utilizamos HRIRs y aplicamos nuestro propio preprocesamiento. Todos los datos utilizados en este trabajo se han adquirido en formato .SOFA y están disponibles en la siguiente dirección web: https://www.sofaconventions.org/mediawiki/index.php/Files .

2.2 Conjuntos de datos utilizados, características y diferencias

En este trabajo, utilizamos datos de 11 conjuntos de datos de HRTF distintos: RIEC [23], FABIAN [24], CIPIC [11], HUTUBS [25], AACHEN [26], LISTEN [27], ARI [28], Crossmod [29], SADIE [30], BiLi [31] y 3D3A [32]. Con el fin de resaltar las principales diferencias entre estos conjuntos de datos que podrían afectar negativamente la compatibilidad de los modelos entrenados con datos diferentes, extrajimos información de diversas fuentes, incluyendo los sitios web de los conjuntos de datos, los documentos asociados y los archivos SOFA.

2.2.1 Condiciones de grabación: cámaras, señales de origen, distancias

Como podemos ver en la Tabla 2, no existe un estándar para la distancia promedio entre las orejas y la distancia a la fuente, lo que puede afectar los niveles de amplitud de las HRTFs resultantes. Además, en los conjuntos de datos que hemos observado, las técnicas derivadas de Swept Sines se utilizan predominantemente para la señal de origen, aunque también se han empleado otros métodos, como ruido pseudo-aleatorio y el Pulso Estirado en el Tiempo Optimizado de Aoshima (OATSP) [33].

Sin embargo, la calidad de las HRTFs resultantes puede verse afectada por las cámaras utilizadas para las grabaciones, que no siempre son anecoicas. Esto puede introducir reflexiones no deseadas y distorsiones que afectan la precisión y confiabilidad de cualquier análisis o modelado subsiguiente basado en estas HRTFs.

Tabla 2 - Diferencias en la configuración de grabación: distancias entre oídos y fuente (en metros), señal de origen y uso de cámara anecoica

Data	Oreja	Fuente	Señal[1]	Anec.
RIEC	0.09	1.5	OATSP	Si
Fabian	0.0662	1.7	Swept S.	Si
CIPIC	0.09	1.5	R. Noise	No [2]
Hutubs	0.75	1.47	M.E.S.S.	Si
AACHEN	0.07	1.2	Swept S.	Semi
Listen	0.09	2.06	Exp. S.S.	Si
ARI	0.09	1.2	Exp. S.S.	Semi
Crossmod	0.09	2.06	Exp. S.S.	Si
SADIE	0.09	1.2	O. Swept S.	Si
BiLi	0.09	2.06	Exp. S.S.	Si
3D3A	0.085	0.76	M.E.S.S.	Si

[1]OATSP: Optimized Aoshima's Time-Stretched Pulse
Swept S: Swept Sine
R. Noise: Pseudo-Aleatory Noise
M.E.S.S.: Multiple Exponential Sine Sweep
[2]Room with absorber

2.2.2 Datos de muestra: duración, tasa de muestreo y número de sujetos

Las variaciones en la tasa de muestreo (SR) y el número total de muestras para cada HRTF son factores significativos a considerar al combinar datos de diferentes conjuntos de datos para el entrenamiento del modelo. Durante el preprocesamiento, es importante tener en cuenta los factores mencionados anteriormente, asegurando que las representaciones de datos finales utilizadas para el entrenamiento del modelo sean coherentes en todos los conjuntos de datos.

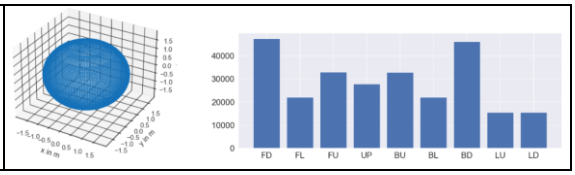
Tabla 3 - Diferencias entre las muestras de distintos conjuntos de Datos: Muestras por HRTF, Frecuencia de muestreo, y número de sujetos.

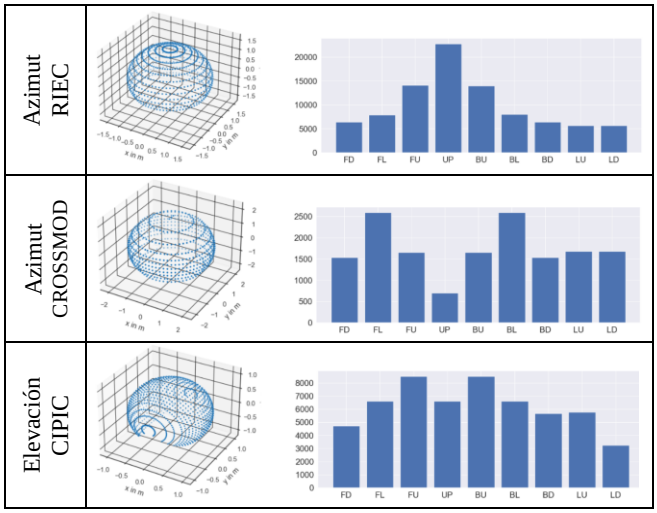
Conjunto Datos	Muestras	SR (kHz)	Nº Suj.
RIEC	512	48	105
Fabian	256	44.1	22
CIPIC	200	44.1	45
Hutubs	256	44.1	96
AACHEN	256	44.1	48
Listen	8192	44.1	50
ARI	256	48	200
Crossmod	8192	44.1	24
SADIE	256	48	20
BiLi	512	96	56
3D3A	2048	96	38

2.2.3 Distribución espacial y de clases

La distribución espacial de las mediciones en términos de los ángulos tomados puede variar significativamente entre conjuntos de datos. Pueden surgir diferencias en la densidad de muestras, el rango de ángulos en elevación y la uniformidad de la distribución de muestras. En algunos conjuntos de datos, existen desequilibrios con una mayor densidad de muestras a lo largo de los ángulos de elevación o de azimut. Además, puede haber diferencias en la distribución de clases de los conjuntos de datos, lo que podría afectar el rendimiento de los modelos de HRTF entrenados con ellos (ver Tabla 4).

Tabla 4 - Coordenadas Espaciales y Balanceo de Clases

Balanceado FABIAN	



2.2.4 Muestras postprocesadas

Además, algunos conjuntos de datos han sido sometidos a un procesamiento posterior antes de su publicación, como se resume en la Tabla 5. Este procesamiento posterior incluye ecualización, cortes de frecuencia, simulaciones numéricas, ventana temporal, calibración de ganancia, compensación de baja frecuencia, extensión de baja frecuencia y ecualización de campo difuso. Combinar datos en bruto y datos procesados puede tener algunos efectos.

Tabla 5 - Resumen del Procesamiento de los Conjuntos de Datos en Archivos SOFA: Datos en Crudo, Ecualización, Corte de Frecuencias, Ventana Temporal, Compensación de Bajas Frecuencias, Calibración de Ganancia, Ecualización de Campo Difuso, Simulación Numérica

Dataset	**Raw**	**Eq**	**FCO**	**TW**	**LFC**	**GC**	**DF Eq**	**NS**
RIEC				✓	✓	✓		
Fabian								✓
CIPIC	✓							
Hutubs			✓					
AACHEN			✓					
Listen	✓							
ARI			✓					
Crossmod	✓							
SADIE				✓	✓	✓	✓	
BiLi		✓		✓				
3D3A		✓						

3. TÉCNICAS DE PREPROCESAMIENTO PROBADAS

Nuestro objetivo fue estandarizar datos de diferentes conjuntos de datos grabados en condiciones variables para mejorar el rendimiento de los modelos entrenados con un tipo de datos en la predicción de la ubicación de HRTFs de diferentes conjuntos de datos. Para llevar a cabo esta tarea, experimentamos con diversas técnicas de preprocesamiento aplicadas a las HRTFs en el dominio de la frecuencia y las utilizamos para entrenar diferentes modelos.

3.1 Normalización

Probamos tres técnicas de normalización: sin normalización, normalización min-max y normalización de Energía Promedio en el Ecuador [34]. En la normalización min-max, cada muestra de HRTF tiene un pico en 1 y una muesca en -1. En la normalización de Energía Promedio en el Ecuador, dividimos cada muestra de HRTF por la energía promedio de HRTF en el ecuador (ángulo de elevación cero).

3.2 Mel warping

Aunque esta técnica no se utilizó para estandarizar los datos, la empleamos para evaluar la respuesta del modelo convirtiendo las HRTFs a la Escala Mel, que es una escala de frecuencia más alineada con la percepción humana. Para lograr esto, dividimos el rango de frecuencias en puntos equidistantes en la Escala Mel y seleccionamos los bins de frecuencia que estaban más cerca de sus respectivas frecuencias en Hz.

3.3 Corte de frecuencia

Realizamos pruebas con diferentes rangos efectivos de frecuencias, incluyendo el rango completo [0-22050 Hz] así como varios rangos restringidos como [20-22050 Hz], [20-16000 Hz], [20-22000 Hz], [50-22050 Hz], [50-16000 Hz], [50-22000 Hz], [500-22050 Hz], [500-16000 Hz] y [500-22000 Hz]. Estas pruebas se realizaron con la comprensión de que algunos conjuntos de datos tienen un rango funcional restringido de frecuencias y algunos pueden tener frecuencias más bajas simuladas o procesadas. Además, dado que en nuestro estudio anterior encontramos señales útiles de elevación por debajo de 5 kHz, queríamos evaluar los resultados cuando las frecuencias por debajo son suprimidas.

3.4 Escala de amplitud

Experimentamos con dos escalas diferentes de HRTF, lineal y logarítmica (log10). La escala lineal mantiene una relación de amplitud constante entre las señales de entrada y salida, mientras que la escala logarítmica (log10) comprime la amplitud de las señales de entrada para producir una salida más uniforme desde el punto de vista perceptual.

4. EXPERIMENTOS

Realizamos extensos experimentos en cada conjunto de datos, utilizando los diversos métodos y parámetros descritos anteriormente. Entrenamos modelos individuales para cada conjunto de datos, así como un modelo combinado con un pequeño porcentaje de muestras de cada conjunto de datos (el 10% de los datos de entrenamiento de los otros conjuntos de datos) y algunos modelos combinados con datos de solo algunos conjuntos de datos seleccionados (el porcentaje variaba según el número de conjuntos de datos seleccionados). Luego, probamos cada modelo con sus propios datos de prueba, así como con datos de prueba de diferentes conjuntos de datos. Nos aseguramos de utilizar sujetos diferentes para el entrenamiento y la prueba, y de mantener la consistencia en el preprocesamiento de datos; por ejemplo, si un modelo se entrenó con datos utilizando una transformación de Mel Warping, probamos ese modelo con diferentes conjuntos de datos con el mismo preprocesamiento. Sin embargo, debido a limitaciones de tiempo y la necesidad de entrenar un gran número de modelos para cada experimento, limitamos el entrenamiento de cada modelo a 100 épocas, con una paciencia de 20 (es decir, se detuvo el entrenamiento si no hubo mejora en el rendimiento durante 20 épocas).

4.1 Resultados sobre la influencia del preprocesamiento

Aunque algunos métodos mostraron mejores resultados de precisión para conjuntos de datos y condiciones específicas, en general, no encontramos ningún parámetro que fuera estadísticamente significativo en comparación con los demás en todos los casos (ver Tabla 6). Realizamos experimentos cambiando solo una variable a la vez mientras manteníamos el resto fijo, y después de analizar los resultados, seleccionamos las pruebas finales con los siguientes parámetros: Normalización de Energía Promedio en el Ecúador, Rango completo de frecuencias (sin corte), sin Mel Warping y escala de amplitud lineal. Estos parámetros se eligieron porque produjeron resultados ligeramente mejores.

Tabla 6 – P-Values para distintos parámetros de: Conjunto de datos usado, Amplitud, Normalización, Mel Warp y Corte de Frecuencias

Data	Ampl	Norm	Mel W.	Freq Cut
2.12e-9	.64	.9	.35	.0503

4.2 Resultados entre conjuntos de datos

Como era de esperar, cada modelo obtuvo buenos resultados cuando se probó con el conjunto de datos al que pertenecen sus muestras de entrenamiento. Sin embargo, obtuvimos resultados de precisión deficientes en conjuntos de datos distintos al utilizado para el entrenamiento. El modelo entrenado con una combinación de distintos conjuntos de datos obtuvo el mejor resultado (Figura 4), a pesar de haber sido entrenado con un número reducido de muestras de cada uno. Es importante señalar que el número limitado de épocas de entrenamiento puede haber afectado los resultados.

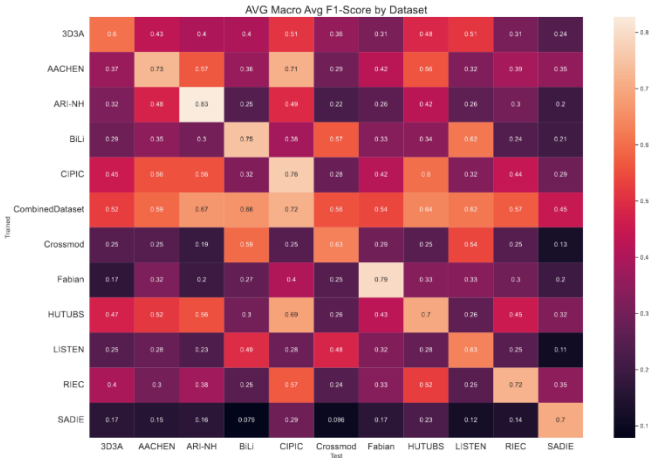


Figura 4 - Resultados de precisión alcanzada por modelos entrenados con un conjunto de datos (filas) contra otros conjuntos de datos distintos (columnas).

5. DISCUSIÓN

Observamos un fenómeno interesante en el que ciertos conjuntos o grupos de conjuntos de datos demostraron una mayor precisión dentro de sí mismos que contra otros conjuntos de datos. Identificamos dos grupos distintos: uno compuesto por los conjuntos de datos Crossmod, BiLi y Listen, y otro que consiste en CIPIC, Riec, Hutubs y AACHEN.

Los conjuntos de datos en el primer grupo, Crossmod, BiLi y Listen, comparten varias similitudes, como la misma distancia entre las orejas (0.09), la misma distancia a la fuente (2.06), el uso de una cámara anecoica y la misma señal de origen (Sine Sweep Exponencial). Por otro lado, los conjuntos de datos en el segundo grupo, CIPIC, Riec, Hutubs y AACHEN, no parecen compartir ninguna característica discernible que explique los resultados de precisión entre conjuntos de datos observados.

6. TRABAJO FUTURO

Nuestro trabajo futuro implica la aplicación de técnicas de Inteligencia Artificial Explicable (XAI) para analizar los resultados obtenidos al trabajar con diversos conjuntos de datos. Nuestro objetivo es identificar los factores responsables del bajo rendimiento de los modelos de clasificación entrenados con datos diferentes y determinar los factores más significativos que contribuyen a este comportamiento. Investigaremos diferentes condiciones que pueden afectar las grabaciones de HRTF, como la distancia entre las orejas, la distancia al emisor, la señal de origen utilizada y las técnicas de procesamiento de datos. Además, analizaremos cómo estas condiciones afectan la prominencia de las bandas de frecuencia de HRTF y su importancia para el modelo de clasificación. Una vez identificadas las causas,

nuestro objetivo es proponer técnicas de estandarización adecuadas para trabajar con conjuntos de datos heterogéneos de HRTF en problemas relacionados, como la personalización de HRTF.

7. AGRADECIMIENTOS

Este trabajo ha sido apoyado por las Becas FPU20/05384 y RYC2020-030679-I del MCIN/AEI/10.13039/501100011033, así como por el "Fondo Social Europeo (FSE) Invertir en tu Futuro". Subvención TED2021-131003B-C21 financiada por MCIN/AEI/10.13039/501100011033 y por el "Fondo de la Unión Europea NextGenerationEU/PRTR". Los autores también agradecen los recursos informáticos Artemisa financiados por el FEDER y la Comunitat Valenciana, y el apoyo técnico del IFIC (CSIC-UV).

8. REFERENCIAS

[1] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *Journal of the Audio Engineering Society*, vol. 43, no. 5, pp. 300–321, 1995.

[2] H. Nakashima, R. Kouyama, N. Hiruma, and Y.-i. Fujisaka, "Binaural wind noise detection, cancellation and its evaluation for hearing aids based on HRTF cues," pp. 004896–004899, 2015.

[3] M. Geronazzo, E. Sikstro¨m, J. Kleimola, F. Avanzini, A. De Go¨tzen, and S. Serafin, "The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 90–97, IEEE, 2018.

[4] M. Zhu, M. Shahnawaz, S. Tubaro, and A. Sarti, "HRTF personalization based on weighted sparse representation of anthropometric features," in *2017 International Conference on 3D Immersion (IC3D)*, pp. 1–7, IEEE, 2017.

[5] G.-T. Lee, S.-M. Choi, B.-Y. Ko, and Y.-H. Park, "HRTF measurement for accurate identification of binaural sound localization cues," *arXiv preprint arXiv:2203.03166*, 2022.

[6] K. Iida and Y. Ishii, "Individualization of the head-related transfer functions on the basis of the spectral cues for sound localization," in *Principles and appli- cations of spatial hearing*, pp. 159–178, World Scien- tific, 2011.

[7] A. Alves-Pinto, A. R. Palmer, and E. A. Lopez-Poveda, "Perception and coding of high-frequency spectral notches: potential implications for sound lo- calization," *Frontiers in neuroscience*, vol. 8, p. 112, 2014.

[8] J. Hebrank and 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, no. 6, pp. 1829–1834, 1974.

[9] R. A. Butler and K. Belendiuk, "Spectral cues utilized in the localization of sound in the median sagittal plane," *The Journal of the Acoustical Society of Amer- ica*, vol. 61, no. 5, pp. 1264–1269, 1977.

[10] J. A. De Rus, A. Lopez-Garc´ıa, J. Lopez-Ballester, J. J. Lopez, A. M. Torres, F. J. Ferri, M. Montagud, and M. Cobos, "On the Application of Explainable Artificial Intelligence Techniques on HRTF Data," in *24th International Congress on Accoustics Proceed- ings*, (Gyeongju, Korea), Oct. 2022.

[11] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pp. 99–102, IEEE, 2001.

[12] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[13] E. Thuillier, H. Gamper, and I. J. Tashev, "Spatial au- dio feature discovery with convolutional neural net- works," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6797–6801, IEEE, 2018.

[14] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," *Applied Sciences*, vol. 10, no. 6, p. 2020, 2020.

[15] T. Kim, J. Lee, and J. Nam, "Sample-level cnn architectures for music auto-tagging using raw waveforms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 366–370, IEEE, 2018

[16] S. Kwon, "A cnn-assisted enhanced audio signal processing for speech emotion recognition," Sensors, vol.

20, no. 1, p. 183, 2019.

[17] S. K. Zielin´ski, P. Antoniuk, H. Lee, and D. Johnson, "Automatic discrimination between front and back ensemble locations in hrtf-convolved binaural recordings of music," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, p. 3, 2022.

[18] A. B. Arrieta, N. D´ıaz-Rodr´ıguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garc´ıa, S. Gil-Lo´pez, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[19] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[21] A. Andreopoulou and A. Roginska, "Towards the creation of a standardized HRTF repository," in *Audio Engineering Society Convention 131*, Audio Engineering Society, 2011.

[22] P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger, *et al.*, "Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions," in *Audio Engineering Society Convention 134*, Audio Engineering Society, 2013.

[23] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acoustical science and technology*, vol. 35, no. 3, pp. 159–165, 2014.

[24] F. Brinkmann, A. Lindau, S. Weinzierl, M. Mu¨ller-Trapet, R. Opdam, M. Vorla¨nder, *et al.*, "A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations," *Journal of the Audio Engineering Society*, vol. 65, no. 10, pp. 841–848, 2017.

[25] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A cross-evaluated database of measured and simulated hrtfs including 3d head meshes, anthropometric features, and headphone impulse responses," *Journal of the Audio Engineering Society*, vol. 67, no. 9, pp. 705–718, 2019.

[26] R. Bomhardt, M. de la Fuente Klein, and J. Fels, "A high-resolution head-related transfer function and three-dimensional ear model database," in *Proceedings of Meetings on Acoustics 172ASA*, vol. 29, p. 050002, Acoustical Society of America, 2016.

[27] O. Warusfel, "Listen hrtf database." `http://recherche.ircam.fr/equipes/salles/listen/`, 2023.

[28] I. fu¨r Schallforschung, "Hrtf-database." `https://www.oeaw.ac.at/en/isf/das-institut/software/hrtf-database`.

[29] "Crossmod hrtfs." `https://sofacoustics.org/data/database/crossmod(hrtf)/`.

[30] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database," *Applied Sciences*, vol. 8, no. 11, p. 2029, 2018.

[31] F. Rugeles Ospina, M. Emerit, and B. F. Katz, "The three-dimensional morphological database for spatial hearing research of the bili project," in *Proceedings of Meetings on Acoustics 169ASA*, vol. 23, p. 050001, Acoustical Society of America, 2015.

[32] R. Sridhar, J. G. Tylka, and E. Choueiri, "A database of head-related transfer functions and morphological measurements," in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.

[33] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *The Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1119–1123, 1995.

[34] Y. Zhang, Y. Wang, and Z. Duan, "HRTF Field: Unifying Measured HRTF Magnitude Representation with Neural Fields," *arXiv preprint arXiv:2210.15196*, 2022.