On the Perceptual Validation of Explainable AI-Based HRTF Saliency for Vertical Sound Localization

1st Juan Antonio De Rus
dept. Informática
Universitat de Valencia
Valencia, Spain
0000-0001-8982-2518

2nd Jesus Lopez-Ballester
dept. Informática
Universitat de Valencia
Valencia, Spain
0000-0002-8212-0342

3rd Mario Montagud
dept. Informática
Universitat de Valencia
Valencia, Spain
0000-0002-2398-1505

4th Francesc J. Ferri
dept. Informática
Universitat de Valencia
Valencia, Spain
0000-0002-1543-3568

5th Maximo Cobos
dept. Informática
Universitat de Valencia
Valencia, Spain
0000-0001-7318-3192

***Abstract*—Head-Related Transfer Functions (HRTFs) play a key role in spatial audio, particularly for vertical sound localization, where interaural time and level differences are negligible. While spectral cues essential for elevation perception have been extensively studied, their precise contributions remain elusive. In this work, we investigate the saliency of spectral cues by leveraging Explainable Artificial Intelligence (XAI) techniques applied to a deep learning model trained for HRTF-based elevation classification. Using saliency maps, we identify and systematically ablate key frequency bands in perceptual experiments to assess their role in human sound localization along the median plane. Our preliminary results suggest that removing high-saliency frequency bands degrades localization accuracy, supporting the model's predictions. However, the degree of impairment varies across conditions, indicating a complex interplay between spectral cues and perceptual processing. These findings highlight the potential of XAI for interpreting spatial hearing models and motivate further investigation into the perceptual significance of HRTF saliency.**

***Index Terms*—Spatial audio, Convolutional neural networks, Explainable AI.**

I. Introduction

Head-Related Impulse Responses (HRIRs) are an essential concept in spatial audio and sound localization. They describe how sound waves are altered as they interact with the human head, ears, and torso [1]. Head-related transfer functions (HRTFs) encapsulate in the frequency domain the frequency-dependent filtering effects of these anatomical structures, which vary from person to person, allowing for the localization of sound sources around the listener [2]. HRTFs are particularly important for determining the position of a sound in three-dimensional space [3], enabling listeners to perceive directionality, such as vertical (elevation) localization [4]. Numerous studies have investigated the spectral cues critical for elevation localization, aiming to identify the frequency bands that contribute to this perceptual task [5]–[8]. Recent research has highlighted specific frequency ranges—400 Hz to 1.2 kHz, 4 to 8 kHz, and 12 to 14 kHz—as particularly relevant [9]. Despite these advancements, a complete understanding of elevation cues remains elusive, posing challenges for applications such as binaural spatial audio simulation and personalization. To tackle this issue, various modeling and prediction techniques have been explored, with deep learning methods emerging as a promising approach for HRTF personalization [10]–[13].

Building on prior research [14]–[16] that leveraged Explainable AI (XAI) to analyze spectral saliency in HRTF-based elevation classification, this study aims to validate those theoretical findings through perceptual experiments. Previous work applied Class Activation Mapping (CAM) techniques to convolutional neural networks (CNNs) trained on large-scale HRTF datasets, revealing that mid-to-high frequency bands (notably between 4-10 kHz) played a crucial role in elevation classification. These findings, while consistent with classical studies on spectral cues, were derived purely from model-driven saliency predictions, leaving open the question of their perceptual relevance for human sound localization. To bridge this gap, we conducted a series of user-based experiments designed to assess the perceptual impact of ablating the frequency bands identified as salient by the model. Objective localization tests were performed, where participants listened to stimuli with systematically removed spectral components corresponding to the model's most relevant frequency regions. By analyzing changes in localization accuracy, we aim to determine whether the saliency patterns observed in neural network models align with human perceptual processes. The results provide preliminary insights into the role of these spectral cues in median-plane sound localization and highlight

the potential of XAI techniques for interpreting and refining computational models of spatial hearing.

II. Background

The ability to localize sound sources in the vertical plane depends on spectral filtering effects introduced by the listener's anatomy. Unlike horizontal localization, where interaural time and level differences (ITD and ILD) provide robust cues, elevation perception relies on monaural spectral cues caused by reflections and diffractions from the pinnae, head, and torso. Early studies identified spectral peaks and notches as key elements in elevation perception. Hebrank and Wright, [5] along with Shaw [6] demonstrated that a spectral notch in the 5-8 kHz range plays a crucial role in front-back and elevation discrimination. Subsequent research highlighted additional peaks at 7-9 kHz and high-frequency cutoffs between 10-14 kHz as relevant for vertical localization. More recent studies [9] have identified three critical bands: 400 Hz - 1.2 kHz, 4 - 8 kHz, and 12 - 14 kHz. Despite these insights, the precise contributions of different frequency bands remain debated. Some studies [17]–[19], suggest that spectral gradients, rather than absolute peaks and notches, may be key cues used by the auditory system, while others [20] emphasize the potential role of low frequencies (below 2 kHz), particularly for front-back discrimination. These complexities have motivated the use of data-driven methods, such as deep learning, to systematically analyze spectral cues in elevation perception.

A. XAI-Based Saliencies

Deep learning models, particularly CNNs, have been employed to classify HRTFs and predict elevation sectors based on spectral patterns. However, the black-box nature of these models raises challenges in understanding how they make decisions. XAI techniques, such as CAM and Grad-CAM, provide a way to visualize the most influential spectral regions for classification. In our previous work [14]–[16], we trained a CNN model for HRTF-based elevation classification, leveraging data from 11 public HRTF datasets to analyze spectral saliencies. These datasets varied in acquisition conditions, including spatial resolution, subject diversity, and measurement techniques. To improve model robustness across datasets, we explored different preprocessing techniques, such as amplitude normalization using *Average Equator Energy* (AEE) [21], mel-frequency warping or Equidistant Rectangular Bandwidth (ERB) filtering. The final trained CNN model used ERB filtering and AEE normalization, which provided the best trade-off between classification accuracy and interpretability. The model classified HRTFs into seven elevation sectors ranging from Front-Down to Back-Down, following a sector-based approach to improve robustness.

Using CAM-based saliency analysis, we identified high-saliency regions primarily in the 4-10 kHz range, with additional contributions from low-frequency bands below 500 Hz in rear and lateral positions. The model consistently highlighted spectral features that align with classical auditory research, such as notches in the 5-8 kHz range and peaks near 7-9 kHz. Additionally, saliencies suggested that low-frequency components may play a greater role in rear localization than previously assumed. To illustrate how the model highlights critical frequency regions, Figure 1 presents an example saliency map for a representative subject from the CIPIC dataset, revealing which frequency bands most influenced the model's classification. While these findings are insightful, they remain model-driven. It is still unclear whether these AI-derived spectral cues correspond to perceptually relevant localization features in human listeners. By combining computational analysis and perceptual validation, this study aims to determine whether XAI-derived spectral cues reflect actual auditory perception.

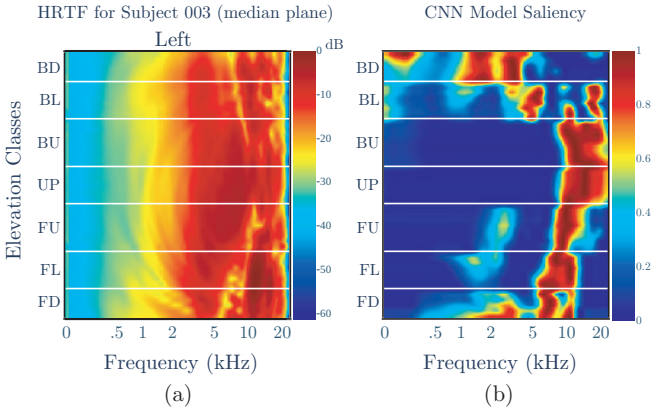


Fig. 1: CAM saliency example on HRTFs from Subject 003 of the CIPIC dataset. (a) HRTF magnitude map divided into the considered elevation sectors. (b) CAM-based saliency.

III. Methodology

A. HRIR Measurement

The experiment was conducted in a controlled recording studio environment, where participants' pinnae were positioned at a height of 1.12 meters. The Presonus Eris E4.5 loudspeakers were used for sound presentation, arranged in a circular configuration with a 1.75-meter radius around the listener to ensure consistent spatial positioning of the stimuli.

For HRIR recordings, we employed Roland CS-10EM in-ear microphones, which provided a high-fidelity capture of individual binaural responses. The same Roland CS-10EM microphones were later used as headphones for stimulus reproduction during the perceptual tests, ensuring consistency between recording and playback conditions. To achieve precise speaker placement and minimize positioning errors, a laser-based positioning system was utilized, allowing for accurate alignment of the loudspeakers relative to the participant. Additionally, all recorded HRTFs were compensated for the frequency response of the microphones, ensuring that the rendered stimuli faithfully represented the individualized acoustic filtering effects of each listener.

A photograph of the measurement setup is shown in Figure 2. The measurement setup included seven speaker po-

sitions along the sagittal plane, selected to assess elevation perception (see Table I). Furthermore, four equatorial positions were placed at 90-degree azimuth separations to provide spatial coverage at 0º elevation that were used to apply AEE normalization.

A total of six subjects (4 males and 2 females) without declared hearing impairment participated in the experiment.

TABLE I: Test locations for elevation perception assessment

Angle in Sagittal Plane	Elevation Class	Abbreviation
-30º	Front-Down	FD
0º	Front-Level	FL
45º	Front-Up	FU
90º	Up	UP
135º	Back-Up	BU
180º	Back-Level	BL
210º	Back-Down	BD

Fig. 2: HRTF measurement setup.

B. HRTFs processing

Once the HRIRs of the subjects are measured, they must be processed into a format compatible with the 1D-CNN model used for elevation classification. This pre-processing step ensures that the input data aligns with the model's expected representation, allowing for CAM analysis to extract the specific frequency regions considered relevant for elevation discrimination. The measured HRIRs are first transformed from the time domain to the frequency domain using the Fast Fourier Transform (FFT). Only the absolute magnitude values are retained, as the model focuses on spectral magnitude features rather than phase information. To standardize the spectral representation and enhance the perceptual alignment of the data, we apply two preprocessing techniques:

a) AEE Normalization [21]: : The amplitude of each HRTF is normalized relative to the mean energy of HRTFs measured at 0° elevation across all azimuths. This ensures consistency by compensating for overall amplitude variations while preserving elevation-dependent spectral features.

b) ERB-Based Spectral Filtering: : The frequency representation is further refined using an ERB filter bank with 255 filters, covering the frequency range 50 Hz to 22.050 Hz [22]. This approach mimics the human cochlear frequency response, enhancing the model's ability to capture perceptually relevant spectral cues.

By processing the HRIRs with AEE normalization and ERB filtering, we generate input data that is directly compatible with the trained 1D-CNN classifier. This enables CAM-based saliency analysis, allowing us to identify the subject-specific frequency regions that the model deems most relevant for elevation classification for each specific subject.

C. Reference Stimuli

The reference stimuli consisted of two Gaussian white-noise bursts, each lasting 1 second, separated by a 250 ms silent interval. To prevent abrupt onset and offset artifacts that could influence perception, each burst was shaped with 20 ms cosine-squared ramps at both the beginning and end. This smoothing helped maintain natural transitions, minimizing unintended spectral or temporal cues.

D. HRTFs Ablation

The saliency values were derived from the classifier model, which processes HRTFs as two-channel inputs, each with 255 frequency components. For each input, the model assigns an elevation class prediction and generates a saliency score for each frequency, indicating its relative contribution to the classification decision. To identify the most critical frequency bands, we extracted saliency peaks for each subject that exceeded a threshold of 0.5 (on a scale from 0 to 1). These peaks represent the most influential frequency components in the HRTF, as determined by the model. Given that the stimulus energy is evenly distributed across frequencies, we implemented a targeted ablation of 25% of the spectral content by selectively removing frequencies around the detected saliency peaks. To achieve this, we applied a 4th-order Butterworth bandpass filter in the time domain, centering each filter at a detected saliency peak. The filter bandwidths were set to evenly distribute the ablation across all identified peaks, ensuring that exactly 25% of the spectral content was removed.

Ablation was performed using two different frequency representations:

- *Linear frequency scale*: where frequency bands are spaced evenly in Hz across the frequency range.
- *ERB scale*: which follows a perceptual frequency distribution that approximates human auditory filtering.

E. Localization Test

Participants first completed a brief training session to familiarize themselves with the spatial characteristics of the stimuli. This consisted of a binaural listening session, where reference stimuli, convolved with the subjects' HRIRs, were presented over headphones. The stimuli covered all measured angles in a predictable sequence, including the different elevations considered and four lateral positions. Before each playback,

participants were explicitly informed of the intended source position. All participants confirmed that they perceived the sources at the expected locations, ensuring they were properly calibrated for the localization test.

Following training, the localization test was conducted, where participants were presented with three different versions of the stimuli in a randomized manner: reference (non-ablated) stimuli, ablated stimuli using a linear frequency scale, and ablated stimuli using an ERB scale. To provide a consistent reference point, each trial began with two bursts of unmodified Front-Level stimuli, followed by the test stimulus, making a total of four noise bursts per presentation. The test included multiple repetitions of each speaker position and ablated stimuli to ensure response reliability.

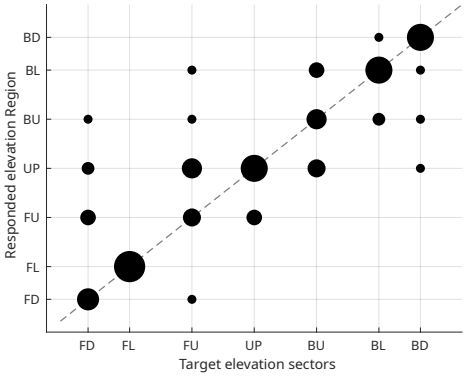
IV. Results

Figure 3 presents three graphs comparing the localization accuracy for unmodified (black) and filtered stimuli (red), where frequency bands were ablated using either a linear or ERB-based approach. The graphs depict the relationship between the participants' identified elevation sectors and the true target elevation sectors. The size of each dot represents the number of participants who selected a given elevation, providing a visual indication of response concentration at each location.

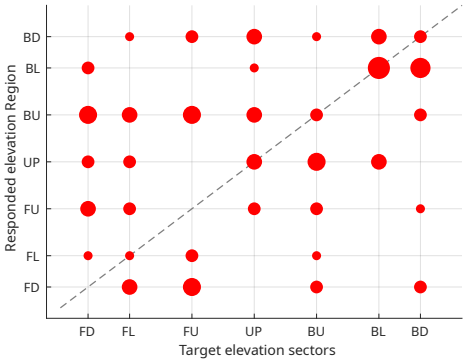
a) Accuracy with Unmodified Stimuli: Fig. 3 a) illustrates the results for the unmodified stimuli, showing a strong concentration of responses along the diagonal, where the user-identified elevation sector matches the true target sector. The presence of larger dots near the diagonal indicates that participants were generally able to accurately identify elevation when presented with full-spectrum stimuli, confirming that no essential spectral cues were missing.

b) Impact of Linear Frequency Filtering: Fig. 3 b) shows results for the linear frequency scale filtered stimuli, where high-saliency frequency bands were removed following a uniform spacing in Hz. The increased dispersion of dots away from the diagonal suggests a higher rate of errors in elevation perception. In particular, the presence of larger dots farther from the diagonal indicates that participants struggled to correctly match the target elevation, suggesting that the filtered frequency bands played a significant role in localization accuracy.

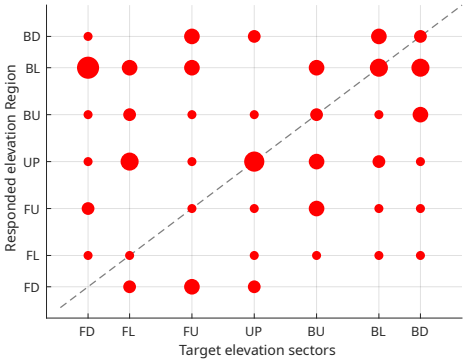
c) Impact of ERB-Scale Filtering: Fig. 3 c) presents the results for the ERB-scale filtered stimuli, where high-saliency frequency bands were removed considering perceptual frequency resolution. The results are similar to those obtained with linear filtering, with increased response variability and more errors in elevation identification compared to the unmodified condition.



(a) Sound localization results using unfiltered stimulus



(b) Sound localization results using filtered stimulus removing 25% of frequency bands on linear scale



(c) Sound localization results using filtered stimulus removing 25% of frequency bands on ERB scale

Fig. 3: Perceived location versus true location for unmodified and filtered stimuli.

V. Discussion

The results of our preliminary listening experiments suggest a correlation between the high-saliency regions predicted by the model and participants' ability to identify elevation. Selective ablation of frequency bands led to a noticeable decrease in localization accuracy, reinforcing the idea that the spectral cues identified by the model are indeed perceptually relevant for human elevation perception. However, due to the limited sample size and participant pool, further experiments are necessary to validate these findings and assess their generalizability.

Our results show that when high-saliency frequency bands—those deemed most critical by the model for elevation classification—were removed, localization accuracy

significantly deteriorated. This confirms that these spectral regions play a crucial role in spatial perception. The results were consistent across both ablation methods (linear frequency scale and ERB-based filtering), with both causing a similar decline in localization accuracy. This suggests that the model-identified high-saliency spectral regions are relevant for elevation perception regardless of the specific frequency scale used for filtering. These findings align with previous research emphasizing the role of frequency-specific information in spatial hearing. Our controlled ablation experiments, guided by XAI methodologies, confirm that the spectral regions identified by our model are fundamental for accurate elevation perception. By selectively removing these bands through saliency-driven ablation, we provide further evidence supporting the importance of spectral cues in HRTFs.

To further investigate the relationship between model-predicted saliencies and human perception, future work will extend the experiments to a larger participant pool. Additionally, we aim to explore different saliency detection thresholds to refine the identification of critical frequency regions and assess the impact of ablation across a wider range of spatial positions and stimuli. These efforts will enhance the generalizability of our findings and provide deeper insights into the perceptual relevance of saliency-based spectral cues.

VI. Conclusions

This study examined the perceptual relevance of high-saliency spectral regions identified by an explainable AI (XAI)-driven model for HRTF-based elevation classification. Through controlled localization experiments, we found that removing model-identified frequency bands led to a decline in elevation accuracy, suggesting that these spectral cues are likely to contribute to human spatial perception.

Both linear and ERB-based ablation resulted in similar impairments, reinforcing the idea that the frequency bands highlighted by the model capture perceptually relevant information. However, given the preliminary nature of this study and the limited participant pool, further validation is necessary to fully establish these findings.

Future work will focus on expanding the participant pool and HRTF datasets, refining saliency detection thresholds, and testing across a wider range of spatial positions and stimuli to further assess the robustness and generalizability of these results.

Acknowledgment

This work has been supported by Grants FPU20/05384 from the Ministry of Universities of Spain, RYC2020-030679-I from MCIN/AEI/-10.13039/501100011033, as well as by the "European Social Fund (ESF) Investing in Your Future". Grant TED2021-131003B-C21 funded by MCIN/AEI/10.13039/501100011033 and by the "EU NextGeneration EU/PRTR". Grant PID2022-137048OB-C41 funded by MICIU/AEI/-10.13039/501100011033 and "ERDF A way of making Europe". The authors acknowledge also the Artemisa computer resources funded by the EU ERDF and Comunitat Valenciana, and the technical support of IFIC (CSIC-UV).

References

[1] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[2] S. Li and J. Peissig, "Measurement of Head-Related Transfer Functions: A Review," *Appl Sci*, vol. 10, no. 14, 2020.

[3] A. Carlini, C. Bordeau, and M. Ambard, "Auditory localization: a comprehensive practical review," *Frontiers in Psychology*, vol. 15, p. 1408073, 2024.

[4] J. Ahveninen, N. Kopčo, and I. P. Jääskeläinen, "Psychophysics and neuronal bases of sound localization in humans," *Hearing Research*, vol. 307, pp. 86–97, 2014.

[5] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *J Acoust Soc Am*, vol. 56, no. 6, pp. 1829–1834, 1974.

[6] E. A. Shaw, "Acoustical features of the human external ear," *Binaural and spatial hearing in real and virtual environments*, vol. 25, p. 47, 1997.

[7] R. A. Butler and K. Belendiuk, "Spectral cues utilized in the localization of sound in the median sagittal plane," *J Acoust Soc Am*, vol. 61, no. 5, pp. 1264–1269, 1977.

[8] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *J Acoust Soc Am*, vol. 112, no. 4, pp. 1583–1596, 09 2002.

[9] D. Yao, J. Li, R. Xia, and Y. Yan, "The role of spectral cues in vertical plane elevation perception," *Acoust Sci Technol*, vol. 41, no. 1, pp. 435–438, 2020.

[10] M. Cobos, J. Ahrens, K. Kowalczyk, and A. Politis, "An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–21, 2022.

[11] G. W. Lee and H. Kim, "Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear," *Appl Sci*, vol. 8, p. 2180, 11 2018.

[12] M. Zhao, Z. Sheng, and Y. Fang, "Magnitude modeling of personalized HRTF based on ear images and anthropometric measurements," *Appl Sci*, vol. 12, no. 16, 2022.

[13] E. Thuillier, H. Gamper, and I. J. Tashev, "Spatial Audio Feature Discovery with Convolutional Neural Networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6797–6801.

[14] J. A. De Rus, A. Lopez-García, 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 Proceedings*, Korea, 2022.

[15] J. A. De Rus, J. Lopez-Ballester, M. Montagud, F. Ferri, J. López, and M. Cobos, "Impact of input preprocessing for HRTF elevation classification over multiple datasets," in *10th Convention of the European Acoustics Association Forum Acusticum 2023*, 01 2023, pp. 2161–2168.

[16] J. A. De Rus, J. Lopez-Ballester, M. Montagud, F. J. Ferri, and M. Cobos, "A Data-Driven Exploration of Elevation Cues in HRTFs: An Explainable AI Perspective Across Multiple Datasets," *arXiv preprint arXiv:2503.11312*, 2025.

[17] L. A. Reiss and E. D. Young, "Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus," *J Neurosci*, vol. 25, no. 14, pp. 3680–3691, 2005.

[18] P. Zakarauskas and M. S. Cynader, "A computational theory of spectral cue localization," *J Acoust Soc Am*, vol. 94, no. 3, pp. 1323–1331, 1993.

[19] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *J Acoust Soc Am*, vol. 136, no. 2, pp. 791–802, 2014.

[20] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," *J Acoust Soc Am*, vol. 88, no. 1, pp. 159–168, 1990.

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

[22] A. Francl and J. H. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nat Hum Behav*, vol. 6, no. 1, pp. 111–133, 2022.