DOI: 10.61782/fa.2025.0354

FORUM ACUSTICUM
EURONOISE 2025

EVALUATING ACOUSTIC CHARACTERISTICS USED BY CONDITIONAL GANS FOR ROOM IMPULSE RESPONSE MODELING

Gonzalo Atienza[1*] **Francisco Pastor-Naranjo**[1] **Daniel de la Prida**[2]
Luis A. Azpicueta-Ruiz[3] **Valery Naranjo**[1] **Gema Piñero**[4]
[1] Institute of Research and Innovation in Bioengineering, Universitat Politècnica de València, Spain
[2] Grupo de Investigación en Acústica Arquitectónica, Universidad Politécnica de Madrid, Spain
[3] Dep. Teoría de la Señal y Comunicaciones, Universidad Carlos III, Madrid, Spain
[4] Institute of Telecommunications and Multimedia Applications (iTEAM),
Universitat Politècnica de València, Spain

ABSTRACT

Conditional Generative Adversarial Networks (CGANs) have previously been used to generate simulated Room Impulse Responses (RIRs) in which the conditional embedding was formed by the room dimensions and reverberation time (RT60), along with the spatial coordinates of the microphone and loudspeaker. In this paper we explore the use of CGANs to model a complete set of real RIRs measured in different rooms. To this end, we propose to increase the number of features used in the embedding vector by including acoustic parameters related to a given RIR such as the source-to-microphone distance and its relation to the critical distance, the direct-to-reverberant ratio (DDR), the early-to-total sound energy ratio (D50), and the clarity index (C50). Our interest is twofold: on the one hand, to evaluate the performance of CGANs in modeling real RIRs by means of comprehensive ablation experiments and, on the other hand, to assess the importance of additional acoustic features on model performance. For the latter purpose, we carried out explainability techniques to identify the most relevant input features on CGAN performance. Results demonstrate the effectiveness of our approach in generating realistic RIRs, providing valuable insights for future research in acoustic modeling techniques.

Keywords: Room Impulse Responses, Deep Learning, GAN, SHapley Additive exPlanations

*__Corresponding author__: gonatsel@upv.edu.es.

1. INTRODUCTION

The rapid growth in immersive audio technologies has significantly transformed how users perceive and interact with audiovisual content. While significant advancements have been made in visual quality, the acoustic experience still holds untapped potential for enhancement, particularly within domestic environments where multi-channel audio systems increasingly replace traditional headphones. Accurate acoustic modeling, especially through Room Impulse Responses (RIRs), is essential for the successful implementation of intelligent audio applications such as immersive 3D audio [1], active noise cancellation [2], or sound zone systems [3].

Recent advances in generative deep learning models, specifically Conditional Generative Adversarial Networks (CGANs) [4], have demonstrated promising results in synthesizing realistic RIRs directly in the time domain. However, the problem of understanding precisely how different acoustic features influence the performance of these generative systems is still open. Our work aims to assess this influence by modeling the RIRs of a set of different rooms through a CGAN, such that introducing a certain embedding as its input, the model output can faithfully infer the corresponding RIR.

Therefore, this paper analyzes the importance of different acoustic features encoded in the embedding for CGAN-based room modeling. We expand upon earlier work [5] by incorporating additional acoustically relevant features such as the distance between the source and the microphone, the value of their direct-to-reverberant ratio (DRR), the energy level, and the room critical distance into the conditioning process. Leveraging explainability techniques, particularly SHapley Additive exPlanations (SHAP) [6], we identify and discuss the most meaningful acoustic parameters, providing insights into their roles and interactions within the model.

2. METHODS

2.1 Dataset

The experiments carried out in this work make use of a comprehensive dataset of measured impulse responses originally introduced by Zhao *et al.* [7]. This dataset is formed by recordings captured across various acoustical environments, allowing a comprehensive evaluation of our Conditional Generative Adversarial Network (CGAN) models.

The data collection employed a configuration consisting of a circular array of 60 loudspeakers and two distinct microphone arrays. The loudspeakers were uniformly distributed along a circle of radius 1.5m. Two microphone arrays were utilized: the first was an 8×8 square grid consisting of 64 microphones, spaced uniformly at 4cm intervals; the second featured a dual-layer circular array with 60 microphones distributed across two concentric circles of radii 12cm and 10cm, respectively.

Measurements were systematically acquired in all the rooms from five distinct zones within the loudspeaker circle: four zones (A, B, C, and D) located along the perimeter of an inner circle of radius 0.4m, and a central zone (E). The dataset comprises a total of $260,400$ measured RIRs of 43,480 samples each, recorded at a sampling frequency of 48kHz, resulting in impulse responses of approximately one-second long. For computational efficiency and model training purposes, RIRs were truncated to the initial 0.128 seconds (approximately $6,144$ samples), long enough to capture the direct sound and early reflections of the RIRs, which are essential for acoustic characterization.

2.2 Fundamentals of CWGAN

When evaluating acoustic characteristics used by Conditional GANs for room impulse response modeling, we follow the methodology employed in previous research [4]: the Conditional Wasserstein Generative Adversarial Network (CWGAN) [8]. CWGAN is a specialized variant of the Generative Adversarial Network (GAN) architecture, designed specifically to generate realistic data conditioned on given inputs or features. It consists of two main neural networks: the generator and the critic. The generator creates new, synthetic data samples guided by the provided conditional information, while the critic evaluates the authenticity and quality of these generated samples by comparing them with actual, real-world data.

A significant distinction between a CWGAN and a traditional GAN is its use of conditional information to guide the generation process, coupled with the Wasserstein distance as its training criterion [9]. The Wasserstein distance serves as a metric to measure differences between the distributions of generated and real data, helping to reduce common training problems observed in standard GANs, such as mode collapse, a situation in which a model repeatedly generates very similar outputs, failing to capture the full diversity of the underlying data distribution.

Moreover, the CWGAN architecture integrates additional techniques such as gradient clipping or gradient penalties [10], ensuring that the critic network behaves well throughout training. This stability allows the generator network to explore more realistic data points effectively.

Table 1. Features included in the initial embedding.

Feature	Description
Listener{X, Y, Z}	Spatial coordinates of the microphone (in cm)
Speaker{X, Y, Z}	Spatial coordinates of the loudspeaker (in cm)
Room{X, Y, Z}	Spatial dimensions of the room (in m)
RT60	Reverberation time (in ms) [12]

2.3 Acoustic Features

In our experimental analysis, we first considered a baseline set of geometric values and one acoustic feature as originally employed in [5, 11]. These features, serving as our benchmark, are described in Tab. 1.

These parameters represent fundamental acoustic and spatial attributes of a given environment, which in turn de-

Table 2. New features included in the embedding.

Feature	Description
Distance	Distance between loudspeaker and microphone (in cm)
Energy dB	Total energy of the RIR (in dB)
Critical Distance	Critical distance of the room (in cm)
DRR	Direct-to-Reverberant Ratio (in dB)
D50	Early-to-total sound energy ratio (in dB)
C50	Clarity Index (in dB)

termine the particular propagation of a sound within that enclosure, establishing a reference performance for subsequent model comparisons.

In the second phase of experiments, we expanded upon the initial benchmark by incorporating additional acoustic characteristics into the embedding vector. These new features are described in Tab. 2. The definition of the new acoustic parameters "Critical Distance", "DRR", "D50" and "C50" can be found in [12]. The goal was to replicate the initial experiment's conditions while assessing the impact and relevance of these newly introduced acoustic parameters.

Finally, a third experimental iteration has been carried out after conducting a feature selection process based on the outcomes of the previous experiments. Upon examination of the performance metrics and contribution of each acoustic feature, we selected the most influential features to include in our final model. This methodical approach ensures the final model utilizes only the most significant and informative characteristics for accurate acoustic modeling.

3. RESULTS

In this study, we use the SHAP methodology [6] to perform an interpretability analysis of our CWGAN model [8]. SHAP is an advanced interpretability framework that quantifies the influence of individual features on a model's predictions. It assigns each feature a value that represents its average marginal contribution to the model's output, computed over all possible combinations of input features. This provides a consistent measure of feature importance, enabling a clear visualization and interpretation of the model's behavior.

To visualize and interpret these contributions, we use SHAP summary plots, which is essential to clarify how these visualizations should be interpreted. In SHAP plots, acoustic features are listed vertically and ordered according to their influence on model performance. The magnitude of each feature's influence is represented by the SHAP values along the horizontal axis.

Negative SHAP values indicate that the corresponding feature contributes to reducing the model's error, measured in NMSE (dB), thus improving prediction accuracy. Conversely, positive SHAP values mean that the feature increases the model's error, which is unfavorable.

Each point in the plot represents a specific embedding instance. The color of these points encodes the original numeric value of the feature for that embedding: red points indicate high feature values, while blue points represent lower feature values.

3.1 SHAP Analysis Across Experimental Stages

As described in Section 2.3, to evaluate how the geometrical and acoustic features influence the performance of our CWGAN to model the RIRs, we performed a series of SHAP analyses across three experimental settings. Each setting corresponds to a specific configuration of the conditioning vector: 1) the benchmark model with the base set of features of Tab. 1, 2) an extended model incorporating the additional acoustic parameters of Tab. 2, and 3) a final model using a reduced and optimized subset of features.

Results of the SHAP analysis for the benchmark model are shown in Fig. 1. As said before, each point represents one of the RIRs of the testset, which is a subset of the training dataset, since the goal of the CWGAN is to accurately model previously seen RIRs. The feature value is represented by the color according to the range shown on the right side of the figure, and the SHAP value represents the impact of the corresponding feature on the model error such that the lower the SHAP value, the lower the model error. Features are ordered by increasing SHAP value, i.e., from the most "beneficial" to the most "damaging" for the model performance.

It can be noted from Fig. 1 that ListenerY and ListenerX are the most influential features. This result aligns with the dataset's structure, which includes a rich diversity of microphone positions, providing the model with substantial spatial variation to learn from. In contrast, the Speak-

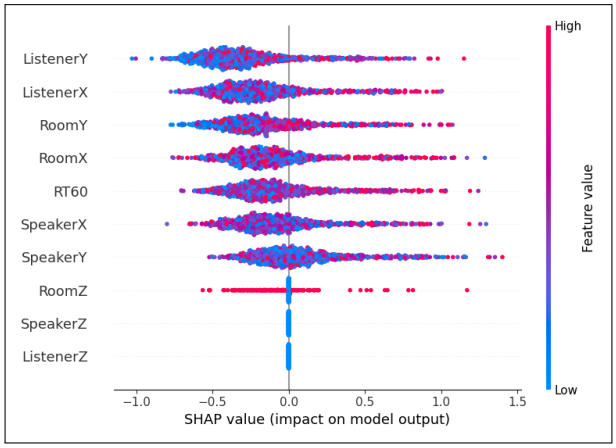

Figure 1. SHAP summary plot for the benchmark model.

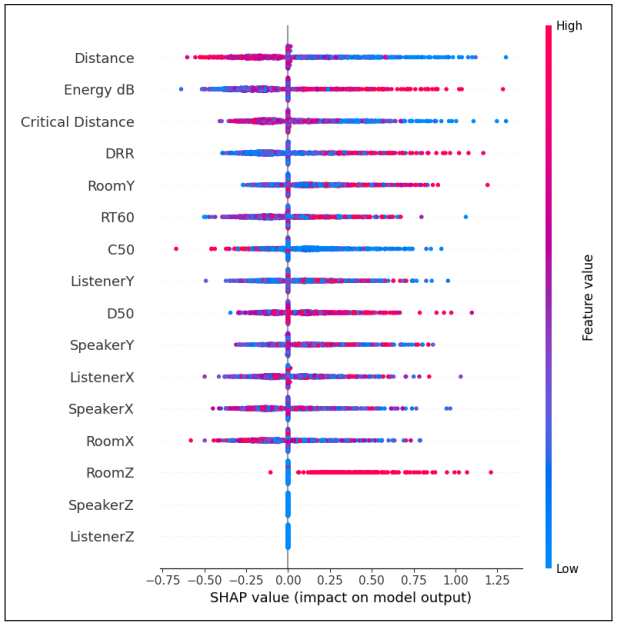

Figure 2. SHAP summary plot for the extended model.

erX and SpeakerY features had a more limited impact, likely due to the smaller number of loudspeakers present in the dataset. RoomX and RoomY, representing the horizontal dimensions of the room, also showed strong influence, in accordance with their acoustic relevance, while RoomZ had little effect, likely because most rooms (except the Hemi-Anechoic chamber) share the same ceiling height [7]. Additionally, SpeakerZ and ListenerZ were among the least influential features, which is expected given that all microphones and loudspeakers are positioned at a fixed height, resulting in no variation along the vertical axis. RT60, in contrast, demonstrated a meaningful contribution, confirming its importance in characterizing reverberant conditions.

In the extended model configuration, shown in Fig. 2, new features from Tab. 2 were introduced to the embedding. Among these, distance emerged as the most relevant feature, with higher distances generally leading to lower model errors. Energy dB also exhibited significant influence, with lower energy levels contributing to better predictions — likely due to the smoother structure of RIRs at lower amplitudes. Although critical distance showed slightly more influence than RT60, the difference was not substantial enough to justify its inclusion in the final model. Given that RT60 is a more widely adopted and representative feature in room acoustics, it was favored during feature selection. Parameters DRR, D50, and C50 revealed different levels of influence, with DRR standing out as the most relevant among them. Since they are closely related, we decided to select the most relevant for the third experiment, the DRR.

A significant difference between results obtained in Fig. 1 and Fig. 2 is that overall SHAP values in this last configuration tended to cluster around zero, indicating that the model had difficulty extracting strong predictive signals from the larger set of features. This suggests that the inclusion of too many features may introduce redundancy and reduce the relative influence of individual parameters.

The final experiment focused on a refined selection of features based on the insights from the previous SHAP analyses, their results are shown in Fig. 3. Compared to the second experiment, a clear improvement in feature relevance is observed: most SHAP values are now shifted toward negative values, indicating that the selected features contribute more consistently to reducing the model error. Additionally, unlike in the benchmark, none of the selected features appear to increase the error on average. This supports the hypothesis that a controlled number of well-chosen representative acoustic characteristics can help the model focus more effectively on relevant information.

To further validate the insights obtained from the SHAP analysis, we evaluated the three models on a held-out test set using two objective metrics: Normalized Mean Squared Error (NMSE) and Normalized Projection Misalignment (NPM) of the RIRs obtained by the CWGAN

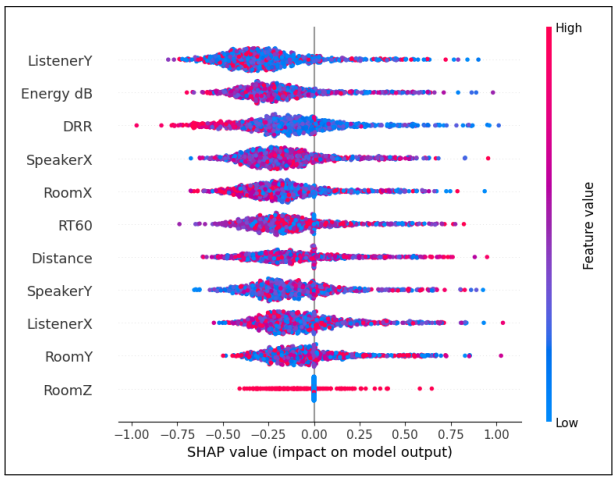

Figure 3. SHAP summary plot for the final model.

model. The NMSE is defined as

$$\text{NMSE} = 10\log_{10}\left(\frac{\|\mathbf{h}(n) - \hat{\mathbf{h}}(n)\|^2}{\|\mathbf{h}(n)\|^2}\right), \qquad (1)$$

where $\mathbf{h}(n)$ is the true time-domain RIR represented as a vector of $6,144$ samples as described in Section 2.1, and $\hat{\mathbf{h}}(n)$ is the CWGAN modeled RIR. The the NPM is defined as

$$\text{NPM} = 10\log_{10}\left(\frac{\|\mathbf{h}(n) - \beta\,\hat{\mathbf{h}}(n)\|^2}{\|\mathbf{h}(n)\|^2}\right), \qquad (2)$$

where

$$\beta = \frac{\mathbf{h}^T(n)\hat{\mathbf{h}}(n)}{\hat{\mathbf{h}}^T(n)\hat{\mathbf{h}}(n)}. \qquad (3)$$

The mean and variance of both metrics are summarized in Tab. 3.

Interestingly, while the SHAP results suggested a more structured and consistent influence of the selected features in the final model, this result did not translate into a lower NMSE or NPM on the test set. In fact, the benchmark model slightly outperformed both the extended and final models in terms of average error. The extended model showed the worst performance, consistent with the SHAP findings where feature influence was diffuse and close to zero. Although the final model improved upon the extended one, it still fell short of the benchmark in terms of raw error.

This discrepancy highlights an important consideration: SHAP provides a valuable lens into feature relevance and model interpretability, but it does not guarantee improvements in performance metrics. The results suggest that, although feature selection helped reduce redundancy and improve focus, the benchmark configuration may still capture additional information beneficial for RIR synthesis.

Table 3. Mean and (variance) in dB of the NMSE and NPM metrics computed over the test set.

Model	NMSE (dB)	NPM (dB)
Benchmark	-2.11 (3.4)	-2.96 (1.5)
Extended	-0.80 (5.6)	-2.34 (1.1)
Final	-0.96 (5.5)	-2.45 (1.2)

4. CONCLUSIONS

This work has presented an interpretability-driven analysis of Conditional GANs for modeling Room Impulse Responses using SHAP explanations. By progressively modifying the feature set across three experimental configurations, we were able to assess not only the predictive relevance of individual acoustic features, but also the effect of their selection on model performance.

Our results show that while the benchmark model, using a base set of spatial and reverberation parameters, achieved the best quantitative performance, the final model demonstrated a more interpretable use of its inputs, with all selected features contributing positively to the precision of the prediction. The extended configuration, which incorporated a large number of features, resulted in degraded performance and limited interpretability, highlighting the risk of over-parameterization.

Ultimately, this study underscores the importance of balancing interpretability with performance in generative acoustic modeling. Explainability tools such as SHAP offer valuable insights that can guide feature selection, improve model transparency, and inform future architectures for RIR synthesis and modeling.

5. ACKNOWLEDGMENTS

This work has been partially funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU through Grant PID2021-124280OB-C21, GVA through Grant 2023-CIPROM/2022/20 and the Spanish Ministry

of Science, Innovation and Universities through Grant FPU23/02726.

6. REFERENCES

[1] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D audio - the new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, 2015.

[2] S. Elliott, *Signal Processing for Active Control*. London: Academic Press, 2001.

[3] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Processing Magazine*, vol. 32, pp. 81–91, 2015.

[4] M. Mirza and S. Osindero, "Conditional generative adversarial nets," in *arXiv preprint arXiv:1411.1784*, 2014.

[5] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "Fast-RIR: Fast neural diffuse room impulse response generator," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 571–575, 2022.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] S. Zhao, Q. Zhu, E. Cheng, and I. S. Burnett, "A room impulse response database for multizone sound field reproduction (L)," *The Journal of the Acoustical Society of America*, vol. 152, pp. 2505–2512, 2022.

[8] C. Fabbri, "Conditional Wasserstein generative adversarial networks," 2017.

[9] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.

[10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028*, 2017.

[11] I. Martin, F. Pastor, F. Fuentes-Hurtado, J. Belloch, L. Azpicueta-Ruiz, V. Naranjo, and G. Piñero, "Predicting room impulse responses through encoder-decoder convolutional neural networks," in *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2023.

[12] H. Kuttruff, *Room Acoustics*. Spon Press, 2009.