Interpretability-Informed Curriculum Learning for Blind Estimation of Room Acoustic Parameters with a Vision Transformer

Luis Paulo Albuquerque Guedes[1,*], Mariane Rembold Petraglia[1] and Julio Cesar Boscher Torres[1]

[1]Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Abstract

Blind estimation of room acoustic parameters from reverberant speech is relevant when the room impulse response is unavailable. This work studies the estimation of reverberation time (RT_{60}) and speech clarity (C_{50}) from log-Mel spectrograms using deep neural networks under a unified protocol. Four architectures are compared under identical conditions: two ResNet-50-based models and two ViT-B/16-based models with different output-head organizations. The strongest baseline is a band-oriented Vision Transformer (M3). Grad-CAM is then used to inspect this model and motivate a deterministic band-based curriculum learning strategy that progressively relaxes frequency masking during training. Here, Grad-CAM is used as a practical interpretability tool to guide training design rather than as evidence of causal importance. Across three datasets of increasing acoustic complexity, the ViT-based models consistently outperform the convolutional baselines, and the curriculum-trained version further improves RT_{60} estimation while providing smaller gains for C_{50}. The main contribution is an interpretability-informed curriculum design for blind acoustic parameter estimation with clearer scope and improved reproducibility.

Keywords

Room acoustics, blind estimation, Vision Transformer, Curriculum Learning, Grad-CAM, Explainable Artificial Intelligence

1. Introduction

Blind estimation of room acoustic parameters from reverberant speech is important when the room impulse response (RIR) is unavailable. Among ISO 3382 descriptors, reverberation time (RT_{60}) and clarity (C_{50}) provide complementary information about room acoustics [1].

Recent learning-based methods have improved blind estimation from time–frequency representations [2, 3]. CNNs are effective for local spectrotemporal modeling [4, 5], whereas transformers capture broader dependencies [6]. Interpretability tools such as Grad-CAM can also support model inspection [7].

This paper does not propose a new explanation method. Instead, it uses interpretability as a practical aid for acoustic regression design. We first compare four architectures for joint multiband estimation of RT_{60} and C_{50} under a unified protocol. We then analyze the best baseline with Grad-CAM and use the observed activation patterns to define a deterministic band-based curriculum learning strategy.

The main contributions are: (i) a controlled comparison of convolutional and transformer-based models for joint multiband estimation of RT_{60} and C_{50}; (ii) a reproducible Grad-CAM-informed curriculum; and (iii) evaluation on datasets of increasing acoustic complexity, including out-of-distribution testing, with public code for reproducibility.

The study is organized in two stages: baseline comparison and curriculum-based optimization. Section 2 reviews the background, Section 3 presents the methodology, Section 4 reports the results, and Section 5 concludes the paper.

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Corresponding author.

luis.guedes@coppe.ufrj.br (L. P. A. Guedes); mariane@poli.ufrj.br (M. R. Petraglia); julio@poli.ufrj.br (J. C. B. Torres)

0000-0002-0978-3123 (L. P. A. Guedes); 0000-0002-9451-3522 (M. R. Petraglia); 0000-0003-2848-3369 (J. C. B. Torres)

2. Background

2.1. Room acoustic descriptors

Under linear time-invariant assumptions, an enclosure is characterized by its room impulse response $h(t)$, from which standard acoustic descriptors are derived [1]. This work focuses on reverberation time and speech clarity.

The reverberation time RT_{60} is the time required for sound energy to decay by 60 dB after excitation stops. In practice, it is commonly approximated from shorter decay intervals:

$$RT_{60} \approx 3T_{20}, \qquad RT_{60} \approx 2T_{30}. \tag{1}$$

The clarity index C_{50} is defined as the ratio between early and late energy using a 50 ms boundary:

$$C_{50} = 10 \log_{10} \left(\frac{\int_0^{50\,\text{ms}} p^2(t)\, dt}{\int_{50\,\text{ms}}^{\infty} p^2(t)\, dt} \right). \tag{2}$$

While RT_{60} describes global decay, C_{50} is more directly related to intelligibility, making them suitable complementary regression targets.

2.2. Analytical room-acoustic models

Classical room acoustics relates reverberation time to geometry and absorption through models such as

$$RT_{60}^{\text{Sabine}} = 0.16\, \frac{V}{S\alpha}, \qquad RT_{60}^{\text{Eyring}} = \frac{0.161\, V}{-S \ln(1-\alpha)}, \qquad RT_{60}^{\text{Millington}} = \frac{0.16\, V}{-\sum_i S_i \ln(1-\alpha_i)}. \tag{3}$$

Although our approach is data-driven, these expressions highlight the physical dependence of reverberation on volume, surface area, and absorption.

2.3. Neural models, curriculum learning, and interpretability

Time–frequency representations are well suited to audio modeling because they preserve both spectral and temporal structure. CNNs capture local patterns effectively [4, 5], whereas transformers model longer-range dependencies through self-attention [6]. In this work, all models use the same 224×224 log-Mel spectrogram and are trained for joint multiband estimation of RT_{60} and C_{50}.

Curriculum Learning (CL) organizes training from easier to harder conditions [8] and has shown benefits in audio tasks [9, 10, 11, 12, 13, 14]. Here, CL is implemented by progressively controlling the spectral support available during training.

Grad-CAM is a gradient-based interpretability method originally proposed for convolutional networks [7]. Its use in transformers is less direct because explanations depend on token representations rather than conventional feature maps [15]. In this study, it is used only as a practical diagnostic tool to inspect the baseline model and motivate the curriculum design.

3. Methodology

The methodology was designed to ensure physical plausibility, diversity, fair comparison, and reproducible target generation. The pipeline comprises absorption modeling, synthetic room and source generation, RIR simulation, computation of RT_{60} and C_{50}, construction of three speech datasets, generation of 224×224 log-Mel spectrograms, training and evaluation of four neural architectures, Grad-CAM analysis of the best baseline, and curriculum-based retraining.

Absorption coefficients are taken from public PTB data [16], using indoor materials across octave bands from 125 Hz to 4000 Hz. Synthetic rooms are sampled within realistic ranges, and each surface is assigned an absorption coefficient α, with reflection magnitude

$$|R| = \sqrt{\max(0, 1-\alpha)}. \tag{4}$$

Speech and noise sources are randomly positioned, and receivers are modeled binaurally with 9 cm interaural spacing and varying orientations. RIRs are generated with GpuRIR [17], filtered into octave bands, and recomposed in post-processing. The target parameters RT_{60} and C_{50} are computed according to ISO 3382 [1].

Speech segments from LibriSpeech [18] are convolved with the simulated RIRs to build three datasets of increasing difficulty: Dataset I with reverberation only, Dataset II with reverberation plus one noise source, and Dataset III with reverberation, five noise types, and real ARNI samples. The signal-to-noise ratio is uniformly sampled between 0 and 20 dB, making Dataset III the most challenging.

All signals are converted into 224×224 log-Mel spectrograms using a 25 ms Hann window, 224 Mel filters, pre-emphasis of 0.97, and log scaling in dB. Four architectures are evaluated: *M1*, a ResNet-50 with band-specific regression heads; *M2*, a ResNet-50 with parameter-specific heads; *M3*, a ViT-B/16 with band-specific heads; and *M4*, a ViT-B/16 with parameter-specific heads. The best baseline is then selected for Grad-CAM analysis and curriculum-based retraining.

3.1. Evaluation metrics

We report MSE, RMSE, MAE, and PCC. MSE penalizes large deviations, RMSE expresses error in the same unit as the target, and MAE measures average absolute deviation. PCC is defined as

$$\mathrm{PCC} = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2}\,\sqrt{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}}. \tag{5}$$

Together, these metrics capture error magnitude, sensitivity to large deviations, and correlation with target variation.

3.2. Grad-CAM for ViT-based regression

Grad-CAM is applied only after M3 is identified as the best baseline. Let y^c denote one target output. For an internal representation F_k, the Grad-CAM weights are

$$\alpha_k = \frac{1}{Z} \sum_i \frac{\partial y^c}{\partial F_{k,i}}, \tag{6}$$

and the saliency map is

$$L^c_{\text{Grad-CAM}} = \mathrm{ReLU}\left(\sum_k \alpha_k F_k \right). \tag{7}$$

For the ViT, Grad-CAM is computed from the last transformer block before the regression head. The class token is discarded, and the patch tokens are reshaped into a 14×14 grid and resized to the spectrogram resolution. Here, Grad-CAM is used only as a qualitative diagnostic to inspect the baseline and support curriculum design.

3.3. From Grad-CAM observations to curriculum design

After training baseline M3, Grad-CAM is applied to a validation subset covering all octave bands and the three datasets. The saliency maps of RT_{60} and C_{50} are inspected in terms of band concentration, compactness, and stability under increasing acoustic difficulty.

Three recurrent patterns motivate the curriculum: low-frequency bands show more diffuse activations, difficult conditions increase off-band responses, and RT_{60} depends on broader spectrotemporal regions than C_{50}, which is more localized. Based on these observations, we adopt a deterministic curriculum that starts with narrower band-centered inputs and progressively restores the full spectrogram, using distinct release schedules for RT_{60} and C_{50}.

3.4. Interpretability-informed curriculum learning

After selecting M3, we define a deterministic band-based curriculum. Let X be the full log-Mel spectrogram and $M_f(\alpha)$ a band mask controlled by $\alpha \in [0, 1]$. The masked input is

$$X_f(\alpha) = X \odot M_f(\alpha). \tag{8}$$

The mixed input is

$$x_{\text{mix}} = \alpha\, x_{\text{band}} + (1 - \alpha)\, x_{\text{full}}, \tag{9}$$

where x_{band} is the masked input and x_{full} is the original spectrogram. Mask expansion follows

$$e = \lfloor (1 - \alpha)\, E_{\max} \rfloor\,. \tag{10}$$

For C_{50}, $\alpha = 1$ is kept until epoch 10 and decays to 0 by epoch 15. For RT_{60}, α decays more gradually from epoch 10 to epoch 30, reflecting the more localized nature of C_{50} and the broader decay behavior of RT_{60}.

These schedules are heuristic and derived from Grad-CAM inspection rather than claimed to be globally optimal. Validation is always performed with $\alpha = 0$, using the full spectrogram.

3.5. Proposed Model

Figure 1 shows the curriculum-trained M3 architecture with the band-specific masking mechanism and progressive training strategy.

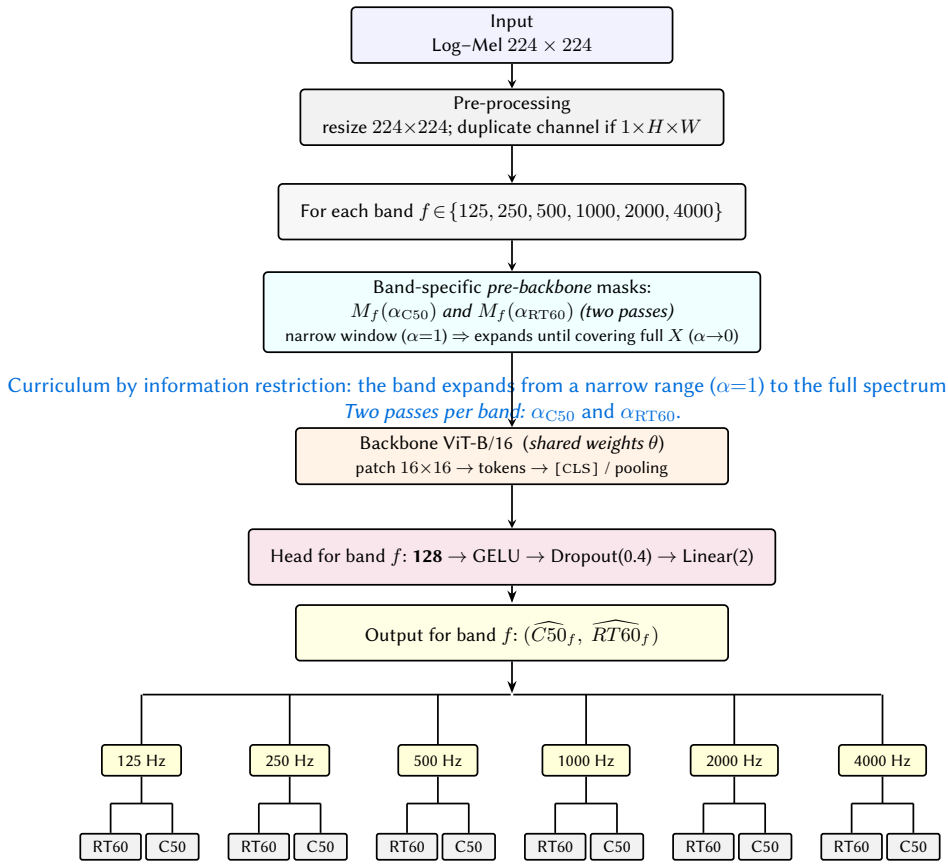


Figure 1: M3 model optimized via *curriculum learning*: information flow (input → pre-processing → band-wise *pre-backbone* masking → ViT *(shared weights)* → band-specific head → outputs $(\widehat{C50}_f, \widehat{RT60}_f)$ for the six frequencies), including target-separated loss, training routine, specialized mask schedule, and mask-free validation.The implementation used in this study is publicly available at: https://github.com/albuca09/AcousticModelling.

4. Results and Discussion

4.1. Comparative evaluation of M1–M4

The first phase compares M1–M4 under identical conditions. Across all datasets, the ViT-based models outperform the ResNet baselines, with M3 achieving the best overall results, particularly for RT_{60}. Performance improves at higher frequencies but decreases from Dataset I to Dataset III as acoustic complexity and the real–synthetic mismatch increase. For brevity, only Dataset III is reported in Table 1, as it represents the most challenging and realistic scenario. Even under these conditions, the curriculum-optimized M3 achieves mean PCC values above 0.98 for RT_{60} and 0.95 for C_{50}.

Table 1
Correlation and error metrics for RT_{60} and C_{50} of M3 with CL on Dataset III.

Metrics	*125 Hz*	*250 Hz*	*500 Hz*	*1000 Hz*	*2000 Hz*	*4000 Hz*	*Mean*
			RT_{60}				
PCC (ρ)	$0.9887 \pm 0.2\%$	$0.9900 \pm 0.1\%$	$0.9898 \pm 0.1\%$	$0.9881 \pm 0.1\%$	$0.9861 \pm 0.1\%$	$0.9833 \pm 0.1\%$	0.9877
MAE (s)	$0.0737 \pm 0.6\%$	$0.0691 \pm 0.6\%$	$0.0716 \pm 0.8\%$	$0.0811 \pm 1.0\%$	$0.0838 \pm 1.1\%$	$0.0903 \pm 1.9\%$	0.0783
MSE (s^2)	$0.0090 \pm 1.0\%$	$0.0078 \pm 0.9\%$	$0.0084 \pm 1.5\%$	$0.0105 \pm 1.8\%$	$0.0111 \pm 2.5\%$	$0.0130 \pm 3.1\%$	0.0100
RMSE (s)	$0.0950 \pm 0.5\%$	$0.0882 \pm 0.5\%$	$0.0914 \pm 0.7\%$	$0.1026 \pm 0.9\%$	$0.1054 \pm 1.3\%$	$0.1139 \pm 1.5\%$	0.0994
MM	$1.0545 \pm 0.2\%$	$1.0519 \pm 0.2\%$	$1.0520 \pm 0.2\%$	$1.0544 \pm 0.2\%$	$1.0486 \pm 0.2\%$	$1.0517 \pm 0.2\%$	1.0522
			C_{50}				
PCC (ρ)	$0.9201 \pm 0.1\%$	$0.9302 \pm 0.2\%$	$0.9506 \pm 0.1\%$	$0.9736 \pm 0.1\%$	$0.9834 \pm 0.1\%$	$0.9856 \pm 0.1\%$	0.9573
MAE (dB)	$1.8429 \pm 0.4\%$	$1.5338 \pm 0.6\%$	$1.3103 \pm 0.4\%$	$1.2153 \pm 0.6\%$	$1.0331 \pm 1.0\%$	$0.9640 \pm 2.3\%$	1.3166
MSE (dB^2)	$5.3294 \pm 0.6\%$	$3.7501 \pm 1.6\%$	$2.7334 \pm 1.7\%$	$2.3486 \pm 2.2\%$	$1.6955 \pm 2.2\%$	$1.4830 \pm 2.4\%$	2.8900
RMSE (dB)	$2.3085 \pm 0.3\%$	$1.9365 \pm 0.8\%$	$1.6533 \pm 0.8\%$	$1.5325 \pm 1.1\%$	$1.3021 \pm 1.1\%$	$1.2178 \pm 1.2\%$	1.6585
MM	$1.6864 \pm 0.5\%$	$1.5795 \pm 0.4\%$	$1.4025 \pm 0.2\%$	$1.3158 \pm 0.2\%$	$1.3045 \pm 0.2\%$	$1.2786 \pm 0.6\%$	1.4279

M3 is selected before curriculum learning, so the second phase compares the original M3 with its curriculum-trained version without confounding model selection and curriculum effects.

4.2. Grad-CAM observations and curriculum learning

Grad-CAM inspection of baseline M3 shows plausible but diffuse attention, especially at lower octave bands and under adverse conditions. This motivates a deterministic band-based curriculum that starts with narrower band-focused inputs and progressively expands to the full spectrogram.

After curriculum learning, M3 improves consistently in RT_{60} and shows smaller but still relevant gains in C_{50}, particularly on the most challenging dataset.

4.3. Trade-off analysis and final results

To analyze the trade-off between performance and temporal context, we evaluate different block sizes. For each RT_{60} range, the smallest block size whose RMSE remains close to the minimum is selected using

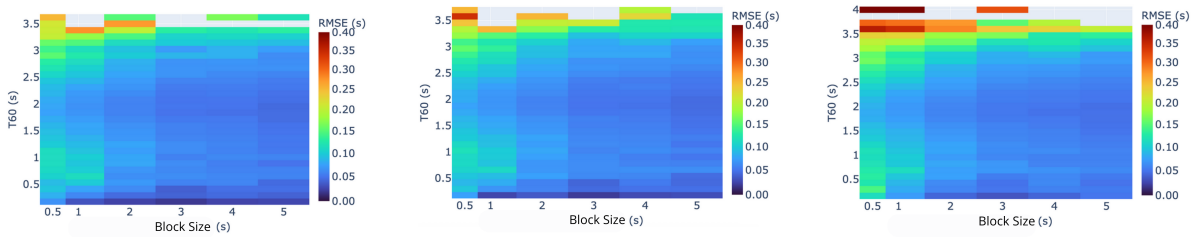
$$\tau = \max\left\{z_{\min} + \varepsilon_{\text{abs}},\ z_{\min}(1 + \varepsilon_{\text{rel}})\right\}. \tag{11}$$

xA global recommendation is then obtained by the mode across RT_{60} ranges. Across datasets, this analysis indicates that $L = 3.0\,\text{s}$ provides a good compromise between error, latency, and computational cost. Figure 2 shows the corresponding RMSE heatmaps for Datasets I–III.

These results support $L = 3.0\,\text{s}$ as the operating point for all subsequent experiments.

Figure 3 is more than illustrative. Before curriculum learning, several bands—especially at low frequencies—show broad or fragmented activations, suggesting diffuse rather than band-coherent evidence. This motivated the progressive band-relaxation strategy.

After curriculum learning, Figure 4 shows more compact and band-consistent saliency with fewer off-band responses, helping explain the observed gains, especially for RT_{60}.



(a) Dataset I (b) Dataset II (c) Dataset III

Figure 2: RMSE heatmaps of RT_{60} as a function of block size and RT_{60} range for the three datasets.

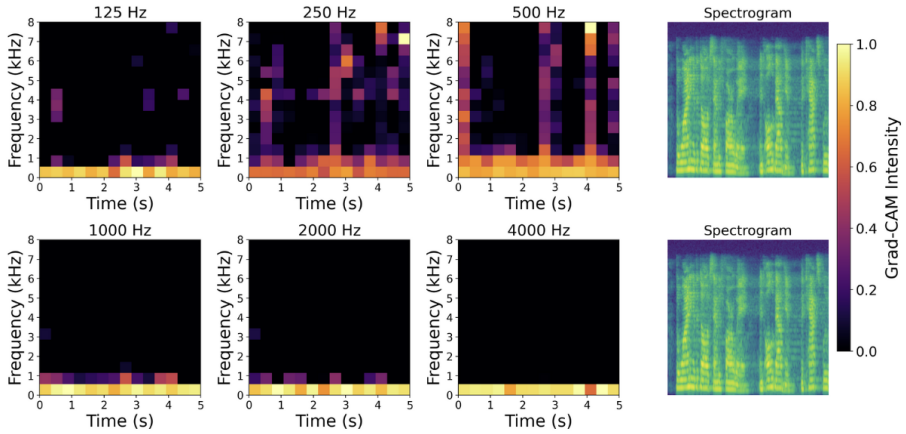


Figure 3: Grad-CAM for the test spectrogram "room_664_src4_rec0_mic1" across six bands for C_{50} before curriculum learning.

Table 2
Global T_{60} (mean across bands) comparison: prior work vs. proposed *M3+CL* ($L = 3\,\mathrm{s}$).

Metric	*Prior works*					*Our work*		
	CNN	*CRNN*	*Pure Att.*	*SS-BRPE*	*SS-BRPE + AUG*	*M3+CL I*	*M3+CL II*	*M3+CL III*
ρ	0.8817	0.9235	0.9660	0.9633	*0.9720*	0.9550	0.9484	0.9209
MAE (s)	0.2966	0.2162	0.1824	0.1470	*0.1312*	0.1357	0.1463	0.1835
MSE (s^2)	0.1473	0.1068	0.0607	0.0479	0.0370	*0.0320*	0.0376	0.0649
MM	1.9952	1.9178	1.4556	1.4029	1.3529	*1.0910*	1.0969	1.1263

Table 3
Performance of the *Optimized M3 (ViT + CL)* across Datasets I–III using $L = 3\,\mathrm{s}$. Results are reported as mean over octave bands (125–4000 Hz).

Metric	*Dataset I*		*Dataset II*		*Dataset III*	
	T_{60}	C_{50}	T_{60}	C_{50}	T_{60}	C_{50}
PCC	0.8712	0.8384	0.8235	0.8011	0.7927	0.7426
RMSE	0.1586	2.2143	0.1749	2.3681	0.1987	2.7452

Table 3 summarizes the mean performance of *Optimized M3 (ViT+CL)* across Datasets I–III, and Table 2 compares global T_{60} results with representative prior work. For global C_{50} estimation, M3 achieves PCC of 0.9549/0.9374 on Datasets II/III, with MAE of 1.3334–1.6443 dB and RMSE of 1.6875–2.3774 dB, remaining competitive with Mel_{128} [24] despite the higher noise complexity of Dataset III.

An additional out-of-distribution evaluation was conducted on the ACE Challenge dataset [23] using *Optimized M3 (ViT+CL)* with $L = 3\,\mathrm{s}$. The results provide a promising external validation with measured room responses, but should be interpreted as partial transfer beyond simulation rather than conclusive

evidence of real-world generalization.

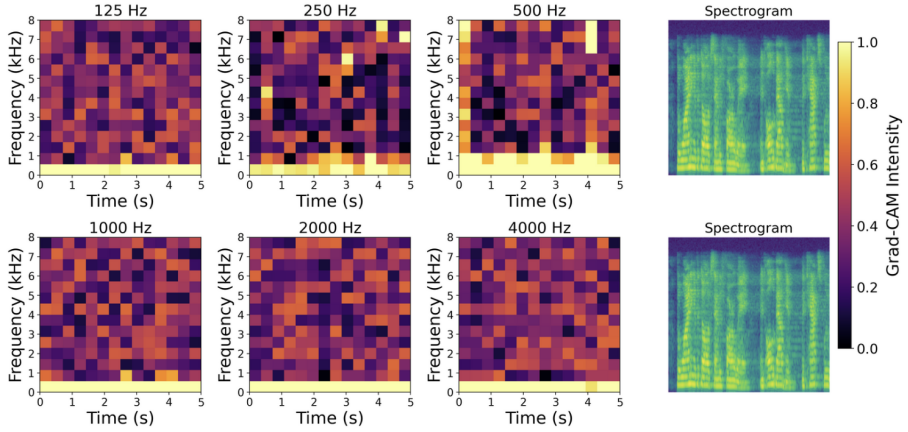


Figure 4: Grad-CAM for the test spectrogram "room_664_src4_rec0_mic1" across six bands for C_{50} after curriculum learning.

On Dataset III with $L = 3\,\mathrm{s}$, *Optimized M3* (ViT-B/16 + CL) requires $126.21\,\mathrm{s}$ for evaluation, $678.06\,\mathrm{MB}$ peak memory, $8.75\,\%$ average CPU, and a $1.61\,\mathrm{GB}$ checkpoint. Overall, M3 provides the best accuracy–consistency trade-off on Dataset III, while curriculum learning improves stability across Datasets I–III and reduces C_{50} error.

5. Conclusion

This paper presents a reproducible study of blind room-acoustic parameter estimation from reverberant speech spectrograms. Among four models, the band-oriented Vision Transformer achieved the best joint estimation of RT_{60} and C_{50}. Grad-CAM revealed frequency-diffuse behavior and motivated a deterministic band-based curriculum that improved performance, particularly for RT_{60}. The contribution is incremental rather than a new XAI framework; future work should include stronger baselines, transformer-specific attribution, faithfulness metrics, and broader real-room validation.

6. Acknowledgments

This work was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) – Finance Code 001.

7. Declarations on Generative AI

No generative AI tools were used in the preparation of this manuscript.

References

[1] International Organization for Standardization: *ISO 3382-1:2009 Acoustics — Measurement of Room Acoustic Parameters — Part 1: Performance Spaces*. ISO, Geneva (2009)

[2] Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press, Cambridge, MA (2016)

[3] Xia, T., et al.: Room acoustic parameter estimation using multi-task learning with convolutional neural networks. *Journal of the Acoustical Society of America* **153**, 1234–1247 (2023)

[4] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (1998)

[5] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. In: *International Conference on Learning Representations (ICLR)* (2021)

[7] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626 (2017)

[8] Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pp. 41–48 (2009)

[9] Gao, T., Du, J., Dai, L.-R., Lee, C.-H.: SNR-based progressive learning of deep neural network for speech enhancement. In: *Proceedings of INTERSPEECH* (2016)

[10] Li, A., Yuan, M., Zheng, C., Li, X.: Speech enhancement using progressive learning-based convolutional recurrent neural network. *Applied Acoustics* **166**, 107347 (2020)

[11] Feng, J., Erol, M.H., Chung, J.S., Senocak, A.: From coarse to fine: Efficient training for audio spectrogram transformers. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2024)

[12] Higuchi, T., Saxena, S., Souden, M., Tran, T.D., Delfarah, M., Dhir, C.: Dynamic curriculum learning via data parameters for noise robust keyword spotting. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021)

[13] Ng, D., Chen, Y., Tian, B., Fu, Q., Chng, E.S.: ConvMixer: Feature interactive convolution with curriculum learning for small footprint and noisy far-field keyword spotting. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2022)

[14] Tonami, K., Imoto, K., Okamoto, T., Yamashita, Y.: Sound event detection based on curriculum learning considering learning difficulty of events. In: *Proceedings of INTERSPEECH* (2021)

[15] Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4190–4197 (2020)

[16] Physikalisch-Technische Bundesanstalt (PTB): Absorption coefficient database. Available at: https://www.ptb.de/cms/ptb/fachabteilungen/abt1/fb-16/ag-163/absorption-coefficient-database.html (Accessed 2025)

[17] Díaz-Guerra, D., Miguel, A., Beltrán, J.R.: GpuRIR: A GPU-based room impulse response generator for interactive auralization. *Multimedia Tools and Applications* **80**, 5653–5678 (2021)

[18] Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: LibriSpeech: An ASR corpus based on public domain audiobooks. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210 (2015)

[19] Ick, C., Mehrabi, A., Jin, W.: Blind acoustic room parameter estimation using phase features. In: *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE (2023)

[20] Wang, C., Jia, M., Li, M., Bao, C., Jin, W.: Attention is all you need for blind room volume estimation. To appear in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024)* (2023)

[21] Wang, C., Jia, M., Li, M., Bao, C., Jin, W.: Exploring the power of pure attention mechanisms in blind room parameter estimation. *EURASIP Journal on Audio, Speech, and Music Processing* **2024**(23) (2024)

[22] Wang, C., Jia, M., Li, M., Bao, C., Jin, W.: SS-BRPE: Self-supervised blind room parameter estimation using attention mechanisms. In: *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (2025)

[23] ACE Challenge: Acoustic Characterization of Environments. Available at: https://www.ace-challenge.org (Accessed 2026)

[24] Gamper, H.: Blind C_{50} estimation from single-channel speech using a convolutional neural network. In: *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, (2020).