

Grad-CAM-Guided Curriculum Learning for Band-Wise Blind Estimation of Room-Acoustic Parameters from Noisy Speech

Luis Paulo Albuquerque Guedes, Mariane Rembold Petraglia and Julio Cesar Boscher Torres

Abstract—This work addresses blind octave-band estimation of reverberation time (T_{60}) and clarity (C_{50}) from noisy reverberant speech. We propose a Grad-CAM-guided curriculum learning strategy for a ViT-based multi-head regressor. Offline Grad-CAM analysis identifies cross-band attention and guides octave-centered time–frequency masks that are progressively relaxed during training. ResNet-50 and ViT-B/16 baselines are evaluated under synthetic and measured RIR conditions. A controlled M3/M3+CL ablation shows the largest gains at low frequencies and more octave-consistent attention. Relative to M3, M3+CL increases the mean PCC by 0.0668 for T_{60} and 0.0199 for C_{50} , including a statistically significant gain of 0.1326 at 125 Hz for T_{60} ($p < 0.001$). At $L = 3.0$ s, M3+CL achieves mean PCC values above 0.98 for T_{60} and 0.95 for C_{50} on Dataset III.

Keywords—blind room-acoustic parameter estimation, reverberation time, clarity, noisy speech, Vision Transformer, Grad-CAM, curriculum learning.

I. INTRODUCTION

Blind estimation of room-acoustic parameters supports dereverberation, speech enhancement, and acoustic monitoring. Reverberation time T_{60} and clarity C_{50} characterize sound decay and early-to-late energy balance, respectively. Their estimation from noisy speech is difficult because speech, room responses, noise, and short-time spectral variations are jointly observed.

Attention-based models, including Vision Transformers (ViTs) [25], have achieved strong room-parameter estimation performance [1]–[3]. However, high predictive accuracy does not ensure that band-wise outputs rely on the corresponding octave regions.

This work proposes a ViT-B/16 multi-head regressor for joint band-wise estimation of T_{60} and C_{50} , trained with a Grad-CAM-guided curriculum learning strategy. Offline Grad-CAM analysis identifies cross-band attention and guides octave-centered masks that are progressively relaxed during training.

The contributions are: (i) joint band-wise estimation with a multi-head ViT; (ii) Grad-CAM-guided curriculum masking; (iii) a controlled ablation against the same architecture without curriculum learning; and (iv) quantitative and qualitative

evaluation of attention consistency, predictive performance, and computational cost under synthetic and measured acoustic conditions.

Section II reviews related work, Section III presents the methodology, Section IV discusses the results, and Section V concludes the paper.

II. RELATED WORK

Classical blind room-acoustic estimators include maximum-likelihood, decay-based, and binaural methods, with the ACE Challenge contributing standardized evaluation protocols [4]. Although efficient and interpretable, these approaches may degrade under noise, non-stationary speech, and short observations [5], [6].

Data-driven methods learn spectro-temporal representations using CNNs, CRNNs, and physically motivated modulation-transfer-function features [7]–[9]. Dedicated C_{50} methods include non-intrusive, joint T_{60} /early-to-late ratio, and CNN–LSTM estimators [23], [27], [28].

Self-supervised and attention-based models further improved spectrogram-based estimation [1]–[3], [10], but generally provide limited evidence that each output uses its target octave band. Grad-CAM [17] and attention rollout [18] can localize influential time–frequency regions. Here, these analyses motivate a curriculum learning strategy [19] that gradually transitions from octave-focused to full-spectrum inputs.

III. MATERIALS AND METHODS

A. Datasets and Acoustic Conditions

Frequency-dependent shoebox RIRs were generated with octave-band absorption coefficients from the PTB database [11]. Reflection coefficients were defined as

$$\beta = \sqrt{1 - \alpha}, \quad (1)$$

where α is the absorption coefficient. RIRs from 125 to 4000 Hz were simulated with `gprRIR` [12] at 16 kHz and recombined into full-band responses.

1) *Speech and Acoustic Scenarios*: LibriSpeech signals [13] with $L \in \{0.5, 1, 2, 3, 4, 5\}$ s were convolved with the RIRs under three conditions: reverberation only, air-conditioning noise, and five simultaneous noises. SNRs ranged from 0 to 20 dB, and channels `_mic1` and `_mic2` were treated as independent receivers.

Three datasets of 50,000 spectrograms were generated. Dataset I contains noise-free reverberant speech. Dataset II

Luis Paulo Albuquerque Guedes, Department of Electrical Engineering, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, e-mail: luis.guedes@coppe.ufrj.br; Mariane Rembold Petraglia, Department of Electrical Engineering, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, e-mail: mariane@poli.ufrj.br; Julio Cesar Boscher Torres, Department of Electrical Engineering, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, e-mail: julio@poli.ufrj.br.

adds air-conditioning noise to samples from 500 Dataset I rooms. Dataset III contains 12,500 samples from each of four conditions: 250 Dataset I rooms, 250 Dataset II rooms, 250 additional rooms with five noises, and measured ARNI rooms [15]. Ground truth for noisy mixtures was averaged across all source–receiver RIRs.

2) *Log-Mel Spectrograms*: Signals were converted into 224×224 log-Mel spectrograms using pre-emphasis of 0.97, a 25-ms Hann window, adaptive hop size, zero padding, and 224 Mel filters from 0 to $f_s/2$. Band-wise T_{60} and C_{50} were computed according to ISO 3382 [20] using `python-acoustics` [21].

B. Partitioning and Experimental Protocol

Each dataset was divided into 80% training, 10% validation, and 10% test sets, corresponding to 40,000, 5,000, and 5,000 samples. Final evaluation used this split.

C. Evaluated Models

Five configurations were evaluated. M1 and M2 use ResNet-50 [26] with six octave-band heads and two target-specific global heads, respectively. M3 and M4 use ViT-B/16 [25] with the same output designs. The proposed **M3+CL** applies curriculum masking to the six-head ViT model M3.

TABLE I: Summary of the evaluated models.

Model	Backbone	Output design	Optimizer	Main role
M1	ResNet-50	Six octave-band heads	Adam	CNN band-wise baseline
M2	ResNet-50	Two global heads	Adam	CNN global-head baseline
M3	ViT-B/16	Six octave-band heads	AdamW	Attention band-wise baseline
M4	ViT-B/16	Two global heads	AdamW	Attention global-head baseline
M3+CL	ViT-B/16	Six octave-band heads	AdamW	Curriculum ablation of M3

Model development used an NVIDIA T4 GPU. Hyperparameters were selected by grid search with five-fold cross-validation and training for up to 100 epochs. All models used batch size 16, random seed 42, and summed MSE loss across octave bands. ResNet-50 models used learning rates of 10^{-5} (backbone) and 10^{-3} (heads) with weight decay 10^{-4} ; ViT-B/16 used 10^{-5} and 3×10^{-4} with weight decay 10^{-2} .

M3 and M3+CL differ only in curriculum masking, isolating the curriculum effect. Table II reports five-fold mean, standard deviation, absolute PCC gain, and 95% CI; all paired comparisons were significant ($p < 0.001$, paired t -test).

D. Grad-CAM-Guided Curriculum Learning

ViT-adapted Grad-CAM and attention rollout were used to analyze time–frequency relevance [17], [18]. Patch-token activations were reshaped into the two-dimensional patch grid, producing output-specific relevance maps. Grad-CAM was computed only for offline analysis and not included in the training loss.

TABLE II: Five-fold cross-validation results for M3 and M3+CL. Values are reported as mean $\pm$ standard deviation.

Band	Target	M3	M3+CL	Gain	95% CI
125 Hz	T_{60}	0.8561 ± 0.0123	0.9887 ± 0.0089	0.1326	[0.1198, 0.1454]
	C_{50}	0.8918 ± 0.0156	0.9201 ± 0.0134	0.0283	[0.0241, 0.0325]
250 Hz	T_{60}	0.8913 ± 0.0112	0.9900 ± 0.0091	0.0987	[0.0889, 0.1085]
	C_{50}	0.9067 ± 0.0145	0.9302 ± 0.0128	0.0235	[0.0198, 0.0272]
500 Hz	T_{60}	0.9216 ± 0.0108	0.9898 ± 0.0087	0.0682	[0.0610, 0.0754]
	C_{50}	0.9282 ± 0.0132	0.9506 ± 0.0119	0.0224	[0.0187, 0.0261]
1000 Hz	T_{60}	0.9443 ± 0.0095	0.9881 ± 0.0083	0.0438	[0.0385, 0.0491]
	C_{50}	0.9553 ± 0.0118	0.9736 ± 0.0104	0.0183	[0.0152, 0.0214]
2000 Hz	T_{60}	0.9522 ± 0.0089	0.9861 ± 0.0078	0.0339	[0.0294, 0.0384]
	C_{50}	0.9704 ± 0.0101	0.9834 ± 0.0092	0.0130	[0.0108, 0.0152]
4000 Hz	T_{60}	0.9597 ± 0.0082	0.9833 ± 0.0074	0.0236	[0.0199, 0.0273]
	C_{50}	0.9720 ± 0.0095	0.9856 ± 0.0087	0.0136	[0.0111, 0.0161]

Because preliminary maps revealed cross-band attention, particularly at low frequencies, M3+CL applies octave-centered masks before the ViT backbone and progressively relaxes them during training. For a log-Mel spectrogram $X \in \mathbb{R}^{F \times T}$ and octave mask $M_f \in \{0, 1\}^{F \times T}$, the masked input is

$$\tilde{X}_f(e) = \alpha(e)M_f \odot X + (1 - \alpha(e))X,$$

where $\alpha(e)$ decays linearly from 1 to 0 over epochs 10 to 30. Training begins with octave-focused inputs and gradually transitions to the full spectrogram. Figure 1 summarizes the M3+CL pipeline.

For quantitative interpretability analysis, the normalized Grad-CAM map G_f is compared with the target octave mask M_f :

$$A_f = \frac{\sum_{i,j} G_f(i,j)M_f(i,j)}{\sum_{i,j} G_f(i,j) + \varepsilon}, \quad (2)$$

where ε prevents division by zero. The score $A_f \in [0, 1]$ represents the fraction of Grad-CAM energy inside the target octave region. It is computed only on the held-out test set and is not part of the training objective.

E. Metrics and Evaluation

Performance is evaluated using the Pearson correlation coefficient (PCC), mean absolute error (MAE), and root mean squared error (RMSE). PCC measures linear agreement, while MAE and RMSE quantify prediction error.

For targets V_n , predictions $\hat{V}_n$, and their sample means, PCC is

$$\rho = \frac{\sum_{n=1}^N (V_n - \bar{V})(\hat{V}_n - \bar{\hat{V}})}{\sqrt{\sum_{n=1}^N (V_n - \bar{V})^2} \sqrt{\sum_{n=1}^N (\hat{V}_n - \bar{\hat{V}})^2}}. \quad (3)$$

The error metrics are

$$\text{MAE} = \frac{1}{N} \sum_{n=1}^N |\hat{V}_n - V_n|, \quad (4)$$

$$\text{MSE} = \frac{1}{N} \sum_{n=1}^N (\hat{V}_n - V_n)^2, \quad (5)$$

and

$$\text{RMSE} = \sqrt{\text{MSE}}. \quad (6)$$

For T_{60} , MAE and RMSE are reported in seconds and MSE in s^2 . For C_{50} , MAE and RMSE are reported in dB and

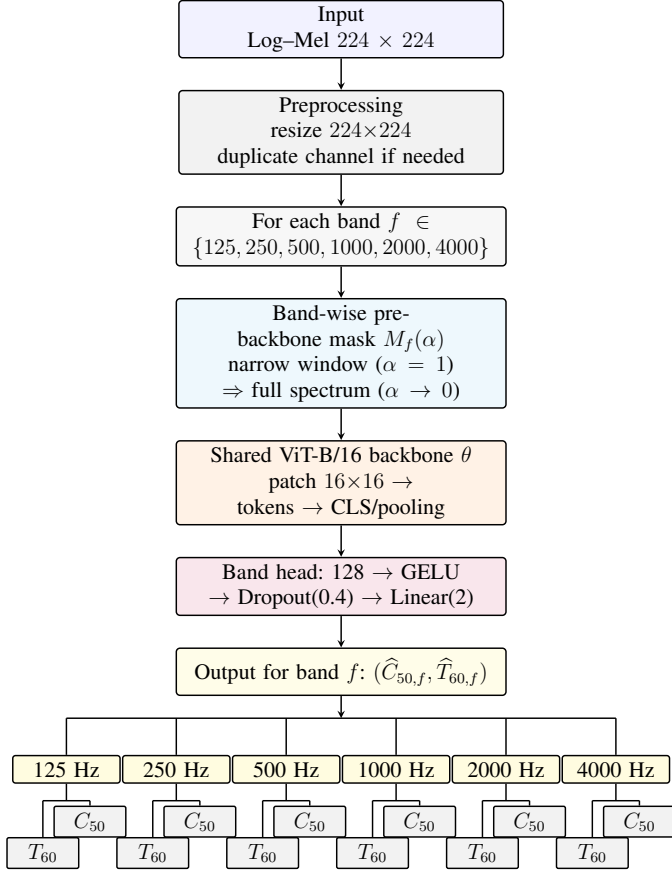


Fig. 1: M3+CL architecture. The model applies a band-wise pre-backbone curriculum mask, processes each masked view with a shared ViT-B/16 backbone, and predicts T_{60} and C_{50} for each octave band.

MSE in dB^2 . MME denotes mean multiplicative error and is reported only for T_{60} .

F. Evaluation Across Acoustic Conditions

Dataset I contains noise-free simulated reverberant speech, Dataset II adds air-conditioning noise, and Dataset III combines multiple noises with measured ARNI RIRs. Because all datasets use the same 80/10/10 split, Dataset III represents increased acoustic complexity rather than strict OOD generalization. Evaluation on ACE [22] is left for future work.

IV. RESULTS AND DISCUSSION

A. Baseline Comparison and Curriculum Ablation

Table III compares M1–M4 and M3+CL on Dataset III. M3 and M3+CL differ only in curriculum masking, isolating its contribution.

M3+CL outperforms M3 across all octave bands, particularly at low frequencies. For T_{60} , the PCC gain decreases from 0.1326 at 125 Hz to 0.0236 at 4000 Hz, with a mean of 0.0668; for C_{50} , the mean gain is 0.0199. At 125 Hz, the T_{60} improvement is statistically significant ($p < 0.001$, paired t -test), with a 95% confidence interval of $[0.1198, 0.1454]$, supporting the contribution of curriculum learning.

TABLE III: PCC per octave band on Dataset III. M3 and M3+CL differ only in the curriculum masking procedure.

Band	Metric	M1	M2	M3	M4	M3+CL
125 Hz	T_{60} PCC	0.815	0.860	0.8561	0.8450	0.9887
	C_{50} PCC	0.873	0.897	0.8918	0.8890	0.9201
250 Hz	T_{60} PCC	0.860	0.860	0.8913	0.8900	0.9900
	C_{50} PCC	0.897	0.899	0.9067	0.9060	0.9302
500 Hz	T_{60} PCC	0.875	0.880	0.9216	0.9201	0.9898
	C_{50} PCC	0.923	0.922	0.9282	0.9280	0.9506
1000 Hz	T_{60} PCC	0.914	0.909	0.9443	0.9420	0.9881
	C_{50} PCC	0.953	0.953	0.9553	0.9552	0.9736
2000 Hz	T_{60} PCC	0.927	0.926	0.9522	0.9520	0.9861
	C_{50} PCC	0.969	0.966	0.9704	0.9703	0.9834
4000 Hz	T_{60} PCC	0.938	0.938	0.9597	0.9596	0.9833
	C_{50} PCC	0.972	0.971	0.9720	0.9719	0.9856

B. Detailed Band-Wise Metrics

Table IV reports the band-wise PCC, MAE, and RMSE achieved by M3+CL across Datasets I–III.

TABLE IV: Band-wise metrics for M3+CL across Datasets I–III. MAE and RMSE are reported in seconds for T_{60} and in dB for C_{50} ; MSE is reported in s^2 and dB^2 , respectively. MME is reported only for T_{60} .

Freq.	Metric	T_{60}			C_{50}		
		I	II	III	I	II	III
125 Hz	PCC	0.9905	0.9897	0.9887	0.9272	0.9237	0.9201
	MAE	0.0663	0.0699	0.0737	1.7267	1.7812	1.8429
	MSE	0.0073	0.0081	0.0090	4.6905	4.9692	5.3294
	RMSE	0.0854	0.0902	0.0950	2.1658	2.2292	2.3085
	MME	1.0493	1.0511	1.0545	–	–	–
250 Hz	PCC	0.9915	0.9908	0.9900	0.9354	0.9329	0.9302
	MAE	0.0624	0.0656	0.0691	1.4550	1.4957	1.5338
	MSE	0.0063	0.0070	0.0078	3.3867	3.5663	3.7501
	RMSE	0.0795	0.0837	0.0882	1.8403	1.8885	1.9365
	MME	1.0474	1.0492	1.0519	–	–	–
500 Hz	PCC	0.9913	0.9906	0.9898	0.9542	0.9522	0.9506
	MAE	0.0645	0.0679	0.0716	1.2470	1.2832	1.3103
	MSE	0.0067	0.0075	0.0084	2.4743	2.6295	2.7334
	RMSE	0.0820	0.0865	0.0914	1.5730	1.6216	1.6533
	MME	1.0479	1.0497	1.0520	–	–	–
1000 Hz	PCC	0.9899	0.9889	0.9881	0.9761	0.9752	0.9736
	MAE	0.0734	0.0773	0.0811	1.1496	1.1810	1.2153
	MSE	0.0087	0.0096	0.0105	2.1042	2.2096	2.3486
	RMSE	0.0931	0.0979	0.1026	1.4506	1.4865	1.5325
	MME	1.0486	1.0509	1.0544	–	–	–
2000 Hz	PCC	0.9885	0.9874	0.9861	0.9846	0.9842	0.9834
	MAE	0.0752	0.0792	0.0838	0.9890	1.0078	1.0331
	MSE	0.0089	0.0099	0.0111	1.5579	1.6152	1.6955
	RMSE	0.0944	0.0996	0.1054	1.2482	1.2709	1.3021
	MME	1.0440	1.0460	1.0486	–	–	–
4000 Hz	PCC	0.9862	0.9850	0.9833	0.9868	0.9862	0.9856
	MAE	0.0808	0.0852	0.0903	0.9171	0.9416	0.9640
	MSE	0.0103	0.0115	0.0130	1.3434	1.4140	1.4830
	RMSE	0.1016	0.1071	0.1139	1.1590	1.1891	1.2178
	MME	1.0458	1.0485	1.0517	–	–	–

Performance remains stable as acoustic complexity increases. For T_{60} , PCC exceeds 0.98 across all bands and datasets, with the lowest errors near 250–500 Hz. C_{50} is more affected by acoustic complexity, particularly at 125 Hz, where modal effects, noise, and short observations impair early-to-late energy estimation.

C. Block-Length Selection

Block duration was selected from RMSE surfaces $Z(x, T_{60})$. Given the minimum error $z_{\min}$, the tolerance was

defined as

$$\tau = \max \{z_{\min} + \varepsilon_{\text{abs}}, z_{\min}(1 + \varepsilon_{\text{rel}})\}. \quad (7)$$

The shortest duration satisfying $\text{RMSE} \leq \tau$ was selected, and the final value was the mode across T_{60} ranges. For Dataset III, $L = 3.0$ s provided the best accuracy–latency trade-off: shorter blocks increased error, whereas longer blocks yielded marginal gains. The same duration was adopted for joint T_{60}/C_{50} estimation, although it was not independently optimized for C_{50} .

D. Comparison with Prior Work

All aggregate results use $L = 3.0$ s. Table V compares global T_{60} performance with prior work, whereas Table IV reports octave-band metrics. Because PCC is nonlinear under pooling, the global PCC does not necessarily equal the mean of the band-wise PCC values.

TABLE V: Global T_{60} comparison: prior work vs. M3+CL using $L = 3.0$ s. Cross-study values are indicative because datasets and protocols differ.

Model	PCC	MAE (s)	MSE (s ²)	RMSE (s)	MME
CNN [24]	0.8817	0.2966	0.1473	0.3838	1.9952
CRNN [1]	0.9235	0.2162	0.1068	0.3268	1.9178
Pure Attention [2]	0.9660	0.1824	0.0607	0.2464	1.4556
SS-BRPE [3]	0.9633	0.1470	0.0479	0.2189	1.4029
SS-BRPE + AUG [3]	0.9720	0.1312	0.0370	0.1924	1.3529
Proposed Model	PCC	MAE (s)	MSE (s ²)	RMSE (s)	MME
M3+CL — Dataset I	0.9550	0.1357	0.0320	0.1789	1.0910
M3+CL — Dataset II	0.9484	0.1463	0.0376	0.1939	1.0969
M3+CL — Dataset III	0.9209	0.1835	0.0649	0.2548	1.1263

M3+CL does not achieve the highest global PCC in the cross-study comparison, but its RMSE at $L = 3.0$ s remains comparable to recent attention-based methods. Because the protocols differ, the comparison is indicative rather than controlled.

Table VI compares C_{50} results. Gamper’s CNN–LSTM reports approximately 2.7 dB RMSE [23], whereas M3+CL values are six-band averages and are not equivalent to broadband estimates.

TABLE VI: Indicative C_{50} comparison with a representative dedicated blind estimator. M3+CL values are means of six band-wise metrics; the CNN–LSTM result follows its original broadband evaluation protocol.

Method	Evaluation condition	ρ	RMSE (dB)
CNN–LSTM Mel128 [23]	Original evaluation set	0.9700	2.68
M3+CL	Dataset I	0.9607	1.5728
M3+CL	Dataset II	0.9591	1.6143
M3+CL	Dataset III	0.9573	1.6585

M3+CL reports a mean band-wise RMSE of 1.6585 dB on Dataset III. Gamper’s model reports a slightly higher PCC, but direct comparison requires evaluation on a common corpus, such as ACE.

E. Aggregate Performance at the Selected Block Length

Table VII summarizes M3+CL at $L = 3.0$ s. Values are averaged across the six octave bands in Table IV, rather than pooled globally.

TABLE VII: Mean band-wise performance of M3+CL across Datasets I–III using $L = 3.0$ s.

Metric	Dataset I		Dataset II		Dataset III	
	T_{60}	C_{50}	T_{60}	C_{50}	T_{60}	C_{50}
PCC	0.9897	0.9607	0.9887	0.9591	0.9877	0.9573
MAE	0.0704	1.2474	0.0742	1.2818	0.0783	1.3166
RMSE	0.0893	1.5728	0.0942	1.6143	0.0994	1.6585

Performance declines slightly from Dataset I to Dataset III, with a larger change for C_{50} because of its sensitivity to noise and early-to-late energy variations.

F. Interpretability Analysis

Figure 2 shows diffuse Grad-CAM attention before curriculum learning, whereas Figure 3 shows more localized, octave-consistent attention.

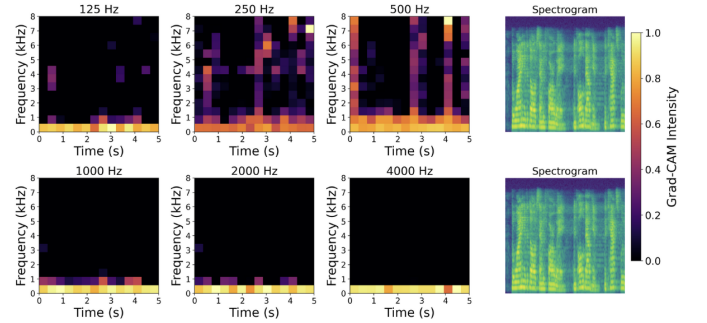


Fig. 2: Grad-CAM maps for C_{50} across six octave bands for the test spectrogram room_664_src4_rec0_mic1, before curriculum learning.

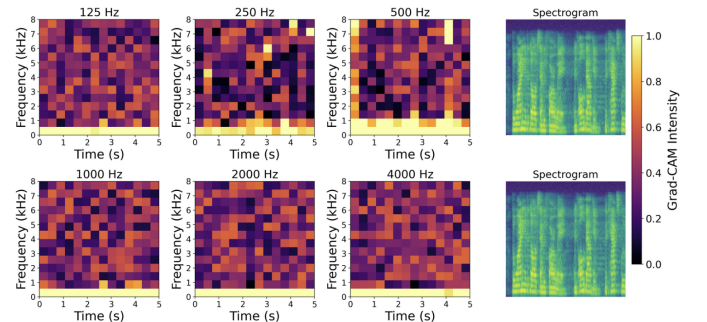


Fig. 3: Grad-CAM maps for C_{50} across six octave bands for the same test spectrogram, after curriculum learning.

The overlap score A_f (Eq. (2)) increased from 0.705 ± 0.067 to 0.925 ± 0.051 ($p < 0.001$, paired t -test), with a 95% confidence interval of $[0.198, 0.242]$ for the mean increase. Gains were largest at 125 Hz (+0.268) and smallest at 4000 Hz (+0.174), consistent with the localization differences in Figures 2 and 3.

G. Computational Cost

Table VIII compares model scale and computational cost under identical hardware and pipeline conditions. M1–M4 values are observed ranges, whereas M3+CL values correspond to the selected configuration.

TABLE VIII: Architecture scale and measured computational requirements.

Model	Backbone params.	Evaluation time (s)	Peak memory (MB)	Checkpoint (GB)
M1	≈ 25.6M	45–55	200–300	0.4–0.6
M2	≈ 25.6M	42–52	195–290	0.4–0.6
M3	≈ 86.6M	115–125	620–660	1.5–1.7
M4	≈ 86.6M	112–122	610–650	1.4–1.6
M3+CL	≈ 86.6M	126.21	678.06	1.61

ResNet-50 models (M1, M2) required less time, memory, and storage than ViT-B/16 models (M3, M4). Differences between M3 and M3+CL reflect run-to-run variability.

For M3+CL on Dataset III ($L = 3.0$ s), evaluating $N_{\text{blocks}} = 1,420$ blocks required 126.21 s, 678.06 MB peak memory, 8.75% average CPU usage, and a 1.61 GB checkpoint, yielding $t_{\text{block}} = 126.21/1,420 = 0.0889$ s/block and $\text{RTF} = 0.0889/3.0 = 0.0296$, or 2.96% of the signal duration.

V. CONCLUSION

This paper proposed M3+CL, a Grad-CAM-guided curriculum strategy for band-wise blind estimation of T_{60} and C_{50} from noisy reverberant speech. The six-head ViT-B/16 model uses progressively relaxed octave-band masks. Relative to M3, its largest PCC gain was 0.1326 for T_{60} at 125 Hz. On Dataset III, mean band-wise PCC reached 0.9877 for T_{60} and 0.9573 for C_{50} . Grad-CAM octave overlap increased from 0.705 ± 0.067 to 0.925 ± 0.051 ($p < 0.001$). Future work will address cross-corpus evaluation on ACE and transfer to non-shoebox and dynamic environments.

REFERENCES

- [1] C. Wang, M. Jia, M. Li, C. Bao and W. Jin, “Attention is all you need for blind room volume estimation,” arXiv:2309.13504, 2023, doi: 10.48550/arXiv.2309.13504.
- [2] C. Wang, M. Jia, M. Li, C. Bao and W. Jin, “Exploring the power of pure attention mechanisms in blind room parameter estimation,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 23, 2024, doi: 10.1186/s13636-024-00344-8.
- [3] C. Wang, M. Jia, M. Li, C. Bao and W. Jin, “SS-BRPE: Self-supervised blind room parameter estimation using attention mechanisms,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10890462.
- [4] J. Eaton, N. D. Gaubitch, A. H. Moore and P. A. Naylor, “The ACE Challenge: Corpus description and performance evaluation,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015, pp. 1–5, doi: 10.1109/WASPAA.2015.7336912.
- [5] R. Ratnam, D. L. Jones and W. D. O’Brien, “Fast algorithms for blind estimation of reverberation time,” *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 537–540, 2004, doi: 10.1109/LSP.2004.826667.
- [6] S. Vesa and A. Härmä, “Automatic estimation of reverberation time from binaural signals,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2005, pp. 281–284, doi: 10.1109/ICASSP.2005.1415701.
- [7] H. Gamper and I. J. Tashev, “Blind reverberation time estimation using a convolutional neural network,” in *Proc. 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 136–140.
- [8] P. Callens and M. Cernak, “Joint blind room acoustic characterization from speech and music signals using convolutional recurrent neural networks,” in *Proc. 24th International Conference on Digital Audio Effects (DAFx-21)*, Vienna, Austria, 2020.
- [9] S. Duangpummet, J. Karnjana, W. Kongprawechnon and M. Unoki, “Blind estimation of room acoustic parameters and speech transmission index using MTF-based CNNs,” in *Proc. 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 181–185, doi: 10.23919/EUSIPCO54536.2021.9616044.
- [10] Y. Gong, C.-I. J. Lai, Y.-A. Chung and J. Glass, “SSAST: Self-supervised audio spectrogram transformer,” in *Proc. AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10699–10709, 2022, doi: 10.1609/aaai.v36i10.21315.
- [11] Physikalisch-Technische Bundesanstalt (PTB), “Absorption coefficient database,” 2025. [Online]. Available: <https://www.ptb.de/cms/ptb/fac-habteilungen/abt1/fb-16/ag-163/absorption-coefficient-database.html>. Accessed: Apr. 24, 2025.
- [12] D. Díaz-Guerra and A. M. Miguel, “gpuRIR: A Python library for room impulse response simulation with GPU acceleration,” arXiv:2010.15271, 2020.
- [13] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, “LibriSpeech: An ASR corpus based on public domain audio books,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [14] A. Mesaros, T. Heittola and T. Virtanen, “Metrics and procedures for the evaluation of sound event detection,” in *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2016.
- [15] Aalto University, “Aalto University room impulse response dataset (ARNI Laboratory).” [Online]. Available: <https://research.aalto.fi/>. Accessed: Jan. 19, 2026.
- [16] J. van Dorp Schuitman, *Auditory Modelling for Assessing Room Acoustics*, Ph.D. dissertation, 2011.
- [17] 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 *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.
- [18] S. Abnar and W. Zuidema, “Quantifying attention flow in transformers,” in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 4190–4197.
- [19] Y. Bengio, J. Louradour, R. Collobert and J. Weston, “Curriculum learning,” in *Proc. 26th Annual International Conference on Machine Learning (ICML)*, 2009, pp. 41–48, doi: 10.1145/1553374.1553380.
- [20] ABNT, *NBR ISO 3382-2: Acústica – Medição de parâmetros de acústica de salas. Parte 2: Tempo de reverberação em salas comuns*, 2017.
- [21] Python-acoustics, “python-acoustics: Python library for acoustics.” [Online]. Available: <https://github.com/python-acoustics/python-acoustics>. Accessed: Jan. 19, 2026.
- [22] ACE Challenge, “Acoustic characterization of environments.” [Online]. Available: <https://www.ace-challenge.org>. Accessed: Jan. 19, 2026.
- [23] H. Gamper, “Blind C50 estimation from single-channel speech using a convolutional neural network,” in *Proc. IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–6.
- [24] C. Ick, A. Mehrabi and W. Jin, “Blind acoustic room parameter estimation using phase features,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
- [26] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [27] P. P. Parada, D. Sharma, J. Lainez, D. Barreda, T. van Waterschoot and P. A. Naylor, “A single-channel non-intrusive C_{50} estimator correlated with speech recognition performance,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 719–732, 2016, doi: 10.1109/TASLP.2015.2507859.
- [28] F. Xiong, S. Goetze, B. Kollmeier and B. T. Meyer, “Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 255–267, 2019, doi: 10.1109/TASLP.2018.2877894.