Effect of noise model mismatch on room-impulse-response representation learning

Karolina Prawda
AudioLab,
School of Physics, Engineering, and Technology,
University of York, York, United Kingdom
karolina.prawda@york.ac.uk

Nils Meyer-Kahlen
Audio Experience Design Team,
Dyson School of Design Engineering,
Imperial College, London, United Kingdom
n.meyerkahlen@imperial.ac.uk

Sebastian J. Schlecht
Multimedia Communications and Signal Processing,
Friedrich-Alexander-Universität Erlangen-Nürnberg,
Erlangen, Germany
sebastian.schlecht@fau.de

***Abstract*—When room impulse responses (RIRs) are used for machine learning (ML) model training, they are often augmented with speech and additive noise. However, if RIRs already include background noise before augmentation, it gets convolved with the speech, creating a noise-model mismatch between the training data and the real recordings of reverberated speech used during inference. This issue is not commonly discussed in audio and acoustic ML model development. This study compares the influence that different noise models arising from the inclusion or exclusion of RIR background noise have on the ML training pipeline. We contrast this against the role of augmentation with additive noise. The results show that using RIRs with truncated background noise leads to compact, separable clusters in the latent space. We also show that models generalize better to realistic reverberated speech if the augmentation during training includes additive noise. The benefits of using truncated background noise in RIRs also translate to reverberation time estimation as a downstream task. The results of this study shed light on the importance of noise model choice for learned RIR representations and inform best practices for creating RIR datasets for ML model training and evaluation using measured RIR data.**

***Index Terms*—room impulse responses, acoustics, machine learning**

I. Introduction

Numerous machine learning (ML) models in acoustics and audio research use room impulse responses (RIRs) in their training pipelines [1], e.g., in artificial reverberation and sound synthesis [2]–[5], RIR estimation [6]–[8] and more. Synthesized RIRs are commonly used [2]–[4], [6] because they enable the fast creation of large, diverse datasets. Measured RIRs are also employed because they include acoustic complexity [2], [7], [9], which is lost when using common synthesis tools, such as the image source method [10], [11]. This paper addresses noise models arising from each of those approaches and their impact on representation learning and on the results of downstream RIR-based tasks.

All authors are also with the Acoustics Lab, Dpt. for Information and Communications Engineering, Aalto University, Espoo, Finland.

Unlike simulated RIRs, measured real-world RIRs contain unavoidable background noise that limits their dynamic range. This noise introduces uncertainty in acoustic parameter estimation [12]. Moreover, when convolved with a source signal, it can lead to an audible effect called "frozen noise" [13], originating from the convolution of the noise floor with the source signal. The issue of background noise and poor signal-to-noise ratio (SNR) is widely discussed in speech enhancement and acoustic echo cancellation research, where masks, or targets, enhance the desired signal and attenuate noise [14]. In room acoustics, those issues are addressed through RIR cropping methods [15], [16], which aim at discarding the response after the point where signal and background noise have the same energy. However, most of the ML literature does not specify whether the RIR background noise is discarded or retained. Only a handful of algorithms explicitly model noise as part of the RIR [8], [17], account for its effect in the loss function [18], or truncate measured RIRs before training [19].

When the RIRs are convolved with source signals, improper truncation of RIRs leads to noise in the training data. During inference on real-world recordings that are not created by convolving source signals with RIRs, this specific kind of noise is absent. The impact of this noise-model-mismatch is unclear. On the one hand, one may assume that using noisy RIR hinders feature learning and generalization to real data. On the other hand, injecting random uncorrelated noise during training is generally thought to act as a form of regularization [20], [21], increasing generalization and robustness [22], [23]. However, this effect is understudied for RIR-based ML tasks, where noisy measured RIRs and clean simulated RIRs are used interchangeably or jointly [2], [24], [25], depending on data availability and the dataset size requirements.

In this study, we systematically analyze how the use of convolutive and additive noise in the training data affects the RIR embeddings in representation learning tasks, focusing on a supervised contrastive learning framework. In the following

section, we explain the differences between the two noise terms in detail. Then, we evaluate whether and how the noise in RIRs affects the discriminability of the latent space and transferability of the learned representations.

II. Problem statement

A typical acoustic measurement consists of an RIR $h[t]$ and a noise term $u[t]$

$$x[t] = h[t] + u[t], \tag{1}$$

where t is discrete time. When such measurements are used in ML training or inference, they are often augmented through convolution with an anechoic signal, such as speech $s[t]$, and another additive noise term $u'[t]$ [7], [9]

$$y^u[t] = s[t] * x[t] + u'[t] = s[t] * (h[t] + u[t]) + u'[t], \tag{2}$$

where $*$ is the convolution operator. Henceforth, $u[t]$ is referred to as *convolutive noise* and $u'[t]$ as *additive noise*. However, such a noise model does not accurately depict a real-world scenario in which signal is recorded in a reverberant room, where convolutive noise is absent

$$y^0[t] = s[t] * h[t] + u'[t]. \tag{3}$$

As a consequence, while training data includes both $u[t]$ and $u'[t]$, real-world data only contains $u'[t]$. The relation in (3) is captured more accurately when noiseless simulated RIRs ($u[t] = 0$) are used for ML tasks instead of measurements, which necessarily have $u[t] > 0$. To the best of the authors' knowledge, the effect of such a noise-model mismatch on ML algorithm results has not been investigated.

Here, we leverage a representation-learning framework in which an upstream encoder maps input signals to compact latent embeddings [6], [7]. We assess how convolutive and additive noise influence latent space structure, whether this effect persists in downstream predictions, and whether noise injection ultimately improves or degrades performance across the full learning pipeline.

III. Methodology

A. Upstream task

Let $\mathcal{H}_0 = \{h_n[t]\}_{n=1}^N$ denote a dataset of clean RIRs and $\mathcal{H}_u = \{h_n[t] + u_n[t]\}_{n=1}^N$ a corresponding dataset of RIRs with convolutive noise. In the upstream task, both datasets are augmented to obtain $y^0[t]$ and $y^u[t]$, respectively, and processed using the same encoder architecture $\mathcal{E}_Y$ to obtain latent representations

$$\begin{aligned} z_n^0 &= \mathcal{E}_Y(y^0[t]), \quad z_n^u = \mathcal{E}_Y(y^u[t]), \\ \mathcal{Z}^0 &= \{z_n^0\}_n, \qquad \mathcal{Z}^u = \{z_n^u\}_n \ . \end{aligned} \tag{4}$$

The effect of noise models on the embeddings is assessed by comparing the latent-space structure using a quality metric $\mathcal{C}(\cdot)$: $\mathcal{C}(\mathcal{Z}^0)$ vs. $\mathcal{C}(\mathcal{Z}^{\mathrm{u}})$.

In the upstream task, we use a contrastive learning approach [26], in which a convolutional neural network encoder $\mathcal{E}_Y$ is trained with a Multi-Positive Contrastive Loss [27]. In this study, $\mathcal{E}_Y$ consists of six convolutional blocks, each comprising a convolutional layer, ReLU activation, and batch normalization; the encoder takes a Mel-spectrogram as input. A subsequent multi-layer perceptron projection head with a single layer and ReLU activation produces the latent representations. The details of the network implementation are described in [6].

B. Downstream task

To reliably assess the impact of convolutive noise as per (2) or lack thereof as in (3) on RIR embeddings, a meaningful downstream task must be chosen. We aim for an approach that does not inherently consider noise (e.g., denoising). As mentioned in Sec. I, some acoustic tasks, such as parameter estimation, are typically noise-sensitive and therefore are well suited to evaluating the noise-mismatch effect. To this end, we consider reverberation time (RT) estimation as the downstream task in this study.

The encoder is kept frozen during the training and evaluation of the downstream network, which contains two fully connected layers and ReLU activation [6]. Unlike in [6], we do not use the additional RT-specific ReLU activation, as the output layer is already linear and extra activation might bias the RT towards values greater than one.

C. Dataset

In this study, we rely exclusively on simulated data to control RIR parameters and noise conditions across the training and evaluation datasets; some experimental conditions require strictly noise-free RIRs. The scope of this study is limited to shoebox rooms generated via the image-source method using *pyroomacoustics* library [11].

We first create the clean RIR dataset, $\mathcal{H}_0$. Following [6], room widths and lengths are uniformly sampled from $[3, 10]$ m, and heights from $[3, 5]$ m. Frequency-dependent absorption coefficients for the six walls are randomly drawn from the built-in *pyroomacoustics* materials library. For the upstream task, we use RIRs from 64 rooms with a single source–receiver position per room. For the downstream task, we use 100 rooms, each with 10 source–receiver positions, for a total of 1000 RIRs. Across both sets, room volumes range from 53–403 m^3, with RTs between 0.2 and 0.9 s.

The dataset of RIRs with convolutive noise, $\mathcal{H}_u$, is created by adding white Gaussian noise, $u_n[t] \sim \mathcal{N}(0, \sigma_u)$, to each RIR in $\mathcal{H}_0$ to achieve peak SNRs between 30–60 dB. To simulate realistic RIR datasets, the convolutive noise $u_n[t]$ is added with different durations: $t_{\mathrm{T}} + \tau$, where t_{T} is the time where the energy of $h_n[t]$ equals the energy of $u_n[t]$ [16], and $\tau = [0, 0.5, 1, 2]$ denotes that $u_n[t]$ extends beyond t_{T} by τ seconds:

$$u_n^\tau[t] = \begin{cases} u_n[t] \sim \mathcal{N}(0, \sigma_u), & t \in [0, t_{\mathrm{T}} + \tau] \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The respective RIR datasets are denoted by $\mathcal{H}_{u^\tau}$.

D. Data augmentation

In the adopted contrastive learning framework, each of the K RIRs in a batch is augmented by convolution with M speech samples $s_{i,m}[t]$, where $i \in 1,\dots,K$ and $m \in 1,\dots,M$. We set $K = 16$ and $M = 48$, and draw speech excerpts $s_{i,m}[t]$ from the LibriSpeech corpus [28]. All RIRs are downsampled to 16 kHz prior to augmentation to match the speech sampling rate. All reverberated speech signals are then truncated to four seconds [6].

Unlike [6], where additive noise $u'_{i,m}[t]$ is included during augmentation (cf. (2) and (3)), we explicitly disentangle the effects of additive noise from those of convolutive noise $u_n[t]$. To this end, we train the encoder $\mathcal{E}_Y$ on each dataset under two conditions: with additive noise augmentation ($u'_{i,m}[t] \sim \mathcal{N}(0,\sigma'_u)$) and without additive noise ($u'_{i,m}[t] = 0$). These conditions are denoted $\mathcal{H}_{u^\tau,\sigma'_u}$ (or $\mathcal{H}_{0,\sigma'_u}$) and $\mathcal{H}_{u^\tau,0}$ (or $\mathcal{H}_{0,0}$), respectively.

E. Evaluation protocols

The effect of the noise model on RIR embeddings is assessed in two stages. First, we evaluate *discriminability*, i.e., the quality of the latent space and the embedding clustering. Afterwards, we assess *transferability*, i.e., downstream task performance [29].

Latent-space clusters are defined by RIR true labels. To quantify discriminability, we compute the following clustering quality metrics $\mathcal{C}(\cdot)$:

- linear classifier on frozen features [29], [30], which quantifies linear separability of clusters; its cross-validation (CV) accuracy over five splits reflects generalization;
- Silhouette (Sil) score, which captures intra-cluster compactness and inter-cluster separation [31], [32]; and
- Davies-Bouldin Index (DBI), which describes variation within clusters and cluster separability [32], [33].

To assess the transferability, we estimate RT from the RIR embeddings and use root mean square error (RMSE), Pearson Correlation Coefficient (PCC), and estimation bias (BIAS) as evaluation metrics.

In both tasks, we compare metrics obtained by running upstream task training and inference (discriminability) and downstream task training (transferability) on in-domain (e.g., $\mathcal{H}_{u^0,0} \rightarrow \mathcal{H}_{u^0,0}$) and out-of-domain data. For the latter, we consider that models should generalize well to realistic data of reverberant speech with additive noise but no convolutive noise, as in (3) (e.g., $\mathcal{H}_{u^0,0} \rightarrow \mathcal{H}_{0,\sigma'_u}$).

IV. Results

This section present the results of discriminability and transferability evaluation on five RIR datasets: the clean RIR dataset $\mathcal{H}_0$, the noisy RIR datasets $\mathcal{H}_{u^\tau}$ with $u_n^\tau[t]$ as per (5), and on two augmentation conditions per each dataset, $u'_{i,m}[t] \sim \mathcal{N}(0,\sigma'_u)$ and $u'_{i,m}[t] = 0$.

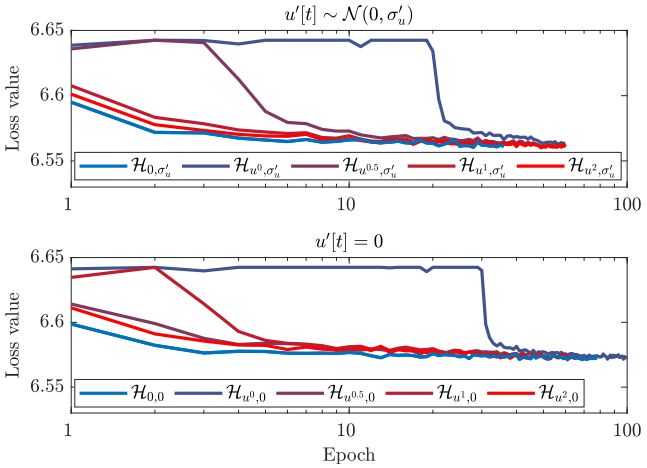

Fig. 1: Loss comparison between RIR datasets with and without convolutive noise. (Top) Augmentation included additive noise. (Bottom) Augmentation did not include additive noise.

A. Loss behavior

Before examining discriminability and transferability, we study upstream training behavior in terms of loss values and convergence rate. Figure 1 shows the loss profiles for the assessed training datasets and augmentation conditions. Including $u'_{i,m}[t]$ accelerates convergence (max. 60 epochs in the top pane of Fig. 1) and reduces the final loss. Models trained on $\mathcal{H}_{0,0}$ and $\mathcal{H}_{0,\sigma'_u}$ start with the lowest loss and converge the fastest. In contrast, training on $\mathcal{H}_{u^0,0}$ and $\mathcal{H}_{u^0,\sigma'_u}$ results in a delayed loss decrease (after 20–30 epochs) and substantially slower convergence, requiring 100 epochs for $\mathcal{H}_{u^0,0}$.

B. Discriminability

The latent space quality metrics in Table I show consistently low Silhouette scores (Sil < 0.25) across all models, even in cases where CV and DBI indicate good cluster separability and compactness, e.g., for $\mathcal{H}_{u^0,0}$. This suggests that the embeddings form neither spherical nor convex clusters [31]. As depicted in the left pane of Fig. 2 for $\mathcal{E}_Y$ trained on $\mathcal{H}_{u^0,0}$, the clusters form tight snake-like structures without overlap, justifying the low Sil scores.

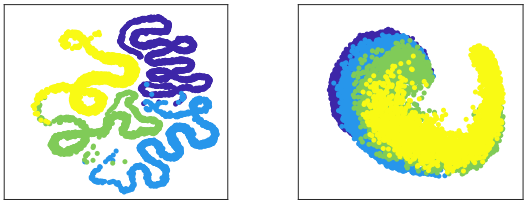

Fig. 2: UMAP projections of embeddings for four RIRs from the $\mathcal{H}_{u^0,0}$ dataset. In-domain inference on $\mathcal{H}_{u^0,0}$ (left) and out-of domain inference on $\mathcal{H}_{0,\sigma'_u}$ (right).

TABLE I: Discriminability results for RIR latent representations on in-domain and out-of-domain data. The best overall models are highlighted.

	In-domain inference			Inference on $\mathcal{H}_{0,\sigma'_u}$		
Train	CV ↑	Sil ↑	DBI ↓	CV ↑	Sil ↑	DBI ↓
$\mathcal{H}_{0,0}$	0.697	0.238	4.65	0.629	-0.140	7.57
$\mathcal{H}_{0,\sigma'_u}$	0.587	0.054	3.15	–	–	–
$\mathcal{H}_{u^0,0}$	**0.758**	**0.201**	**1.68**	0.595	-0.202	6.90
$\mathcal{H}_{u^0,\sigma'_u}$	0.722	0.128	2.08	**0.791**	**0.135**	**2.89**
$\mathcal{H}_{u^{0.5},0}$	0.874	0.036	2.46	0.676	-0.134	7.10
$\mathcal{H}_{u^{0.5},\sigma'_u}$	0.651	0.014	4.17	0.692	-0.004	4.51
$\mathcal{H}_{u^1,0}$	0.781	0.050	4.73	0.547	-0.103	9.86
$\mathcal{H}_{u^1,\sigma'_u}$	0.665	0.015	3.01	0.630	-0.073	5.73
$\mathcal{H}_{u^2,0}$	0.629	0.064	3.54	0.460	-0.106	12.18
$\mathcal{H}_{u^2,\sigma'_u}$	0.643	0.059	3.73	0.516	-0.175	6.95

Using clean RIRs yields lower latent-space quality than when some convolutive noise is used ($\tau \leq 0.5$), indicating a beneficial regularizing effect of convolutive noise. However, when $u_n^\tau[t]$ extends beyond the effective RIR duration ($\tau \geq 1$, bottom rows of Table I), this benefit diminishes and all metrics suggest more diffuse, less separable clusters.

Under realistic inference conditions without convolutive noise ($\mathcal{H}_{0,\sigma'_u}$), latent-space quality degrades for all models except $\mathcal{H}_{u^0,\sigma'_u}$. The degradation is strongest for models trained without additive noise, where Sil < 0 indicates poor clustering and sample misalignment, while DBI > 5 suggests diffuse clusters and small inter-cluster distances. This effect is depicted in Fig. 2 for embeddings learned from $\mathcal{H}_{u^0,0}$: while in-domain inference (left) produces tight, separable yet non-convex clusters, out-of-domain inference on $\mathcal{H}_{0,\sigma'_u}$ (right) leads to more dispersed and overlapping structures. These results indicate that additive noise augmentation is essential for achieving robust generalization to real recordings.

C. Transferability

The transferability results for the RT estimation task are summarized in Table II. The model trained on realistic reverberated speech, $\mathcal{H}_{0,\sigma'_u}$, performs the best of all, followed by $\mathcal{H}_{u^{0.5},0}$ for in-domain and $\mathcal{H}_{u^0,\sigma'_u}$ for out-of domain scenarios, respectively. Among the evaluated metrics, PCC is the most discriminative, separating the high-performing models (PCC > 0.8) from those that produce RT estimates with only weak linear correspondence to ground-truth RT (PCC < 0.4).

These transferability results indicate that having some convolutive noise in the training RIR dataset improves downstream performance, provided that the noise is truncated at or near the energy crossing point t_T. As with the discriminability results, augmentation with additive noise improves generalization to realistic reverberated speech.

V. Discussion

The study shows the impact of noise model on learning RIR representations, though several limitations should be noted. The dataset comprised synthesized RIRs from simple shoebox geometries to enable quick data generation and control over convolutive and additive noise properties. This neglected the complexities of real-world measured RIRs. In addition, white noise was used to represent both additive and convolutive noise, which does not capture colored noise characteristics and frequency-dependent SNR. Finally, only a single network architecture was tested. Although the literature suggests that different architectures can lead to learning similar structures [34], [35], such a comparison is non-trivial [36]. Future work should address these limitations by using measured RIRs, more realistic noise signals, and additional architectures.

TABLE II: Transferability results for RIR latent representations on in-domain and out-of-domain data. The best overall models are highlighted.

(a) In-domain inference.

Upstream	RMSE ↓	PCC ↑	BIAS → 0
$\mathcal{H}_{0,0}$	0.10	0.70	-0.030
$\mathcal{H}_{0,\sigma'_u}$	**0.07**	**0.83**	**0.003**
$\mathcal{H}_{u^0,0}$	0.10	0.67	0.002
$\mathcal{H}_{u^0,\sigma'_u}$	0.11	0.54	-0.003
$\mathcal{H}_{u^{0.5},0}$	**0.07**	**0.83**	**-0.003**
$\mathcal{H}_{u^{0.5},\sigma'_u}$	0.07	0.84	-0.029
$\mathcal{H}_{u^1,0}$	0.11	0.24	0.001
$\mathcal{H}_{u^1,\sigma'_u}$	0.11	0.52	-0.028
$\mathcal{H}_{u^2,0}$	0.12	0.38	0.001
$\mathcal{H}_{u^2,\sigma'_u}$	0.11	0.38	0.001

(b) Inference on the realistic reverberated speech $\mathcal{H}_{0,\sigma'_u}$.

Upstream	RMSE ↓	PCC ↑	BIAS → 0
$\mathcal{H}_{0,0}$	0.09	0.72	-0.009
$\mathcal{H}_{u^0,0}$	0.12	0.60	-0.048
$\mathcal{H}_{u^0,\sigma'_u}$	**0.07**	**0.83**	**-0.014**
$\mathcal{H}_{u^{0.5},0}$	0.09	0.75	-0.029
$\mathcal{H}_{u^{0.5},\sigma'_u}$	0.08	0.82	-0.014
$\mathcal{H}_{u^1,0}$	0.08	0.78	0.003
$\mathcal{H}_{u^1,\sigma'_u}$	0.09	0.82	-0.047
$\mathcal{H}_{u^2,0}$	0.08	0.78	0.001
$\mathcal{H}_{u^2,\sigma'_u}$	0.08	0.78	0.0

VI. Conclusions

This study examined the roles of convolutive and additive noise in learning RIR representations. We show that while the presence of convolutive noise in the training RIRs slows encoder convergence, it improves latent space organization and downstream task performance. Moreover, truncating convolutive noise yields better latent representations and downstream performance on realistic data than allowing the noise floor to extend beyond the effective RIR duration. This emphasizes the importance of appropriate RIR truncation as a pre-processing step in RIR-based ML tasks and in dataset creation.

Our results further demonstrate that, within the scope of the evaluated setup, including additive noise during training improves generalization to real recordings of reverberated speech, which naturally do not contain convolutive noise. Additive noise enhances the latent space quality and the downstream task performance. This highlights the significance of noise-model matching in audio and acoustic ML tasks.

ACKNOWLEDGMENTS

This work was conducted on the Viking cluster, a high-performance computing facility provided by the University of York. The authors thank Philipp Götz for his support in preparing the machine learning pipeline for this study.

REFERENCES

[1] T. Van Waterschoot, "Deep, data-driven modeling of room acoustics: literature review and research perspectives," in *Proceedings of the 11th Convention of the European Acoustics Association Forum Acusticum / EuroNoise 2025*, Málaga, Spain, Dec. 2025, pp. 4065–4072, European Acoustics Association.

[2] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," in *Advances in Neural Information Processing Systems*, 2022, pp. 3165–3177.

[3] K. Su, M. Chen, and E. Shlizerman, "INRAS: Implicit Neural Representation for Audio Scenes," in *Advances in Neural Information Processing Systems*, 2022, pp. 8144–8158.

[4] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu, "Neural acoustic context field: Rendering realistic room impulse response with neural fields," 2023.

[5] G. Dal Santo, B. Alary, K. Prawda, S. Schlecht, and V. Välimäki, "RIR2FDN: an improved room impulse response analysis and synthesis," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*. University of Surrey, 2024, pp. 230–237.

[6] P. Götz, C. Tuna, A. Walther, and E. A. P. Habets, "Contrastive representation learning for acoustic parameter estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[7] P. Götz, C. Tuna, A. Brendel, A. Walther, and E. A. P. Habets, "Blind acoustic parameter estimation through task-agnostic embeddings using latent approximations," in *18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 289–293.

[8] G. Götz, R. Falcón Pérez, S. J. Schlecht, and V. Pulkki, "Neural network for multi-exponential sound energy decay analysis," *J. Acoust. Soc. Am.*, vol. 152, no. 2, pp. 942–953, Aug. 2022.

[9] P. Götz, G. D. Santo, S. J. Schlecht, V. Välimäki, and E. A. P. Habets, "Matching reverberant speech through learned acoustic embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 14577–14581.

[10] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943 – 950, 1979.

[11] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 351—-355.

[12] M. Guski and M. Vorländer, "Comparison of noise compensation methods for room acoustic impulse response evaluations," *Acta Acustica united with Acustica*, vol. 100, no. 2, pp. 320–327, 2014.

[13] P. Massé, T. Carpentier, O. Warusfel, and M. Noisternig, "A robust denoising process for spatial room impulse responses with diffuse reverberation tails," *J. Acoust. Soc. Am.*, vol. 147, no. 4, pp. 2250–2260, 04 2020.

[14] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[15] A. Lundeby, T. E. Vigran, H. Bietz, and M. Vorländer, "Uncertainties of measurements in room acoustics," *Acta Acust. un. Acust.*, vol. 81, no. 4, pp. 344–355, 1995.

[16] K. Prawda, N. Meyer-Kahlen, and S. J. Schlecht, "Cropping room impulse responses using unimodal regression of their covariance," *JASA Express Letters*, vol. 5, no. 8, pp. 081601, 08 2025.

[17] G. Dal Santo, K. Prawda, S. J. Schlecht, and V. Välimäki, "Learning recursive attenuation filters under noisy conditions," *arXiv e-prints*, pp. arXiv–2512, 2025.

[18] A. Richard, P. Dodds, and V. K. Ithapu, "Deep impulse responses: Estimating and parameterizing filters with deep networks," 2022.

[19] A. I. Mezza, R. Giampiccolo, E. De Sena, and A. Bernardini, "Data-driven room acoustic modeling via differentiable feedback delay networks with learnable delay lines," *EURASIP J. Audio Speech Music Process.*, vol. 2024, no. 51, 2024.

[20] C. M. Bishop, "Training with noise is equivalent to Tikhonov regularization," *Neural computation*, vol. 7, no. 1, pp. 108–116, 1995.

[21] O. Dhifallah and Y. Lu, "On the inherent regularization effects of noise injection during training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2665–2675.

[22] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2021, pp. 676–680.

[23] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[24] L. Delebecque and R. Serizel, "BinauRec: A dataset to test the influence of the use of room impulse responses on binaural speech enhancement," in *31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 126–130.

[25] L. Bahrman, M. Fontaine, J. Le Roux, and G. Richard, "Speech dereverberation constrained on room impulse response characteristics," in *Proc. INTERSPEECH*, 2024.

[26] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 18661–18673.

[27] Y. Tian, L. Fan, P. Isola, H. Chang, and D. Krishnan, "StableRep: Synthetic images from text-to-image models make strong visual representation learners," in *Advances in Neural Information Processing Systems*, 2023, vol. 36, pp. 48382–48402.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[29] J.-H. Lee, D. Yoon, B. Ji, K. Kim, and S. Hwang, "Rethinking evaluation protocols of visual representations learned via self-supervised learning," *arXiv preprint arXiv:2304.03456*, 2023.

[30] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[31] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[32] D. Chicco, A. Campagner, A. Spagnolo, D. Ciucci, and G. Jurman, "The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic for unsupervised clustering internal evaluation of two convex clusters," *PeerJ Computer Science*, vol. 11, pp. e3309, Nov. 2025.

[33] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[34] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *International conference on machine learning*. PMlR, 2019, pp. 3519–3529.

[35] A. Morcos, M. Raghu, and S. Bengio, "Insights on representational similarity in neural networks with canonical correlation," in *Advances in Neural Information Processing Systems*, Montréal, Canada, 2018, vol. 31.

[36] P. M. Stecher, V. Sikimic, and R. Kahle, "Do Different Artificial Neural Networks Think Alike? On the Feasibility of Cataloging Representations," *Minds and Machines*, vol. 36, no. 1, pp. 17, 2026.