

Quantifying the Uncertainty of Blindly Estimated Room Embeddings Using a Dispersion-Calibrated Score

Yang Xiang¹, Philipp Götz², Emanuël A. P. Habets², Andreas Walther³, Wenwu Wang¹, Philip J. B. Jackson¹

¹ Centre for Vision Speech and Signal Processing, University of Surrey, United Kingdom

² International Audio Laboratories Erlangen*, Erlangen, Germany

³ Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

yang.xiang@surrey.ac.uk, p.jackson@surrey.ac.uk

Abstract

Room embeddings derived from reverberant speech are often unreliable: speech content and recording degradation can alter the representation even when speaker, room, and source–receiver geometry remain unchanged, degrading downstream task performance. We propose a framework that learns room embeddings robust to speech-content variation and a representation-level uncertainty score from reverberant speech without downstream-task supervision. The embedding is anchored to a structured room impulse response (RIR) latent space and trained using a multi-view data structure with Kullback–Leibler (KL)-based alignment; a multi-positive contrastive term further refines robustness. A lightweight uncertainty head is calibrated using the dispersion of corruption-induced embeddings and optimized with a rank-based objective. Across waveform- and spectrogram-level corruptions, the score is consistent with representation dispersion and enables effective selective prediction while requiring only a single utterance at inference.

Index Terms: room acoustics, reverberant speech, representation learning, contrastive learning, uncertainty estimation

1. Introduction

Reverberant speech carries cues about the acoustic environment, enabling applications such as blind estimation of physical parameters (e.g., reverberation time T_{60} and clarity index C_{50}) [1–3], acoustic-environment retrieval/verification [4] and robust speech processing [5]. Existing approaches broadly fall into two directions: (i) task-specific blind inference, including room impulse response (RIR) reconstruction [6–8] and acoustic-parameter estimation, and (ii) task-agnostic representation learning that yields reusable room embeddings [9–11], which can be transferred across downstream objectives.

However, learning *reliable* room embeddings remains challenging because the observed signal is confounded by non-room factors (e.g., content, speaker, noise). In practice, recordings are rarely clean: noise and partial signal loss (e.g., dropouts, bandwidth limits) can substantially alter embeddings even when the room and source–receiver geometry are unchanged, degrading representation quality and downstream robustness. This motivates a *general-purpose* uncertainty score that indicates whether an embedding should be trusted. While uncertainty estimation has been widely studied in classification and regression [12, 13], recent blind room-acoustic inference work has begun to quantify uncertainty by calibrating error bounds

for RIR/parameter estimates derived from learned environment representations [11]. Yet, predicting task-agnostic uncertainty scores for *representation reliability* with controlled speech corruptions is underexplored.

Following [11], we learn a task-agnostic room representation from reverberant speech together with an uncertainty score reflecting representation reliability under corruption. Our approach has three stages. First, we learn a structured room-acoustic latent space using a variational autoencoder (VAE) trained on log-mel RIR spectrograms. Second, we train a speech encoder with multi-view training and contrastive learning to improve robustness to speech-content variation while remaining aligned with the RIR latent space. Third, we train a lightweight uncertainty head that estimates representation uncertainty from embedding dispersion under controlled corruptions. Compared with [11], Stage-1 uses the same RIR-VAE latent space, while Stage-2 adds multi-view KL-based alignment with a multi-positive contrastive term, and Stage-3 converts corruption-induced dispersion into a task-agnostic, single-utterance uncertainty score.

Our main contributions are: (1) We identify the multi-view data structure (multiple speech realizations per RIR) as a key driver of robustness to speech-content variation under latent alignment. (2) We propose a general-purpose room embedding framework that combines KL-based alignment to a frozen RIR-VAE latent space with a multi-positive contrastive term for a modest verification gain. (3) We introduce a dispersion-calibrated uncertainty score, trained with a rank-based objective to map corruption-induced embedding dispersion to a single-utterance uncertainty estimate.

We evaluate representation quality on RIR verification, RIR log-mel reconstruction, and acoustic-parameter prediction, and assess reliability using uncertainty–dispersion correlation and selective prediction on verification and reconstruction. Overall, the score detects unreliable embeddings under the considered corruptions.

2. Problem Formulation

We model the observed reverberant speech signal $y[t]$ as an anechoic speech source $x[t]$ convolved with an RIR $h[\ell]$ of length L , corrupted by additive noise $w[t]$:

$$y[t] = \sum_{\ell=0}^{L-1} h[\ell] x[t-\ell] + w[t], \quad (1)$$

where t is the discrete-time sample index and ℓ indexes the RIR taps. In this work, both reverberant speech and RIRs are repre-

*A joint institution of Fraunhofer IIS and Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany.

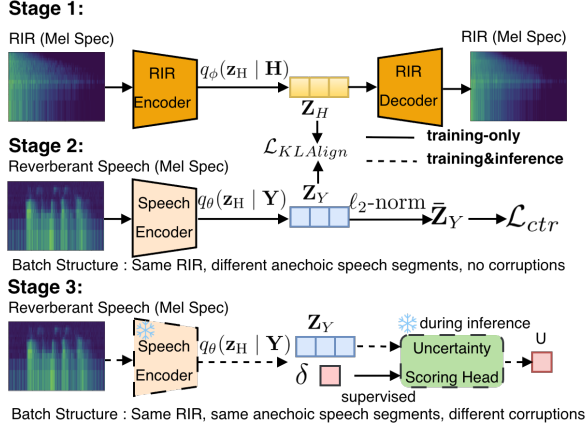


Figure 1: Overview of the three-stage training pipeline. Stage-1 trains an RIR-VAE to obtain posterior statistics $(\mu_{z_H}, \sigma_{z_H}^2)$. Stage-2 learns a speech embedding $\mathbf{z}_Y$ using multi-positive contrastive learning and KL alignment to the frozen RIR posterior. Stage-3 freezes the speech encoder and trains a scalar uncertainty head supervised by embedding dispersion δ .

sented by log-mel spectrograms $\mathbf{Y} \in \mathbb{R}^{F \times T}$ and $\mathbf{H} \in \mathbb{R}^{F \times T}$, with F frequency bands and T time frames.

Given $\mathbf{Y}$ only (blind setting), our goal is to learn a task-agnostic room representation $\mathbf{z}_Y = f_\theta(\mathbf{Y})$ that captures room acoustics while being robust to speech-content variation, and to estimate a scalar uncertainty $U = g_\psi(\mathbf{z}_Y)$ that reflects representation reliability under varying acoustic conditions.

3. Method

Figure 1 summarizes our three-stage training pipeline. Stages 1–2 follow the latent-approximation anchoring strategy in [11], while we introduce (i) multi-positive contrastive learning to improve robustness to speech-content variation and (ii) a dispersion-supervised uncertainty head for predicting task-agnostic representation reliability.

3.1. Stage-1: RIR-VAE Pretraining

We train a VAE [14] on RIR log-mel spectrograms $\mathbf{H} \in \mathbb{R}^{F \times T}$ to obtain a structured latent space. The encoder outputs a diagonal-Gaussian posterior $q_\phi(\mathbf{z}_H | \mathbf{H}) = \mathcal{N}(\mu_{z_H}, \text{diag}(\sigma_{z_H}^2))$, where μ_{z_H} and $\sigma_{z_H}^2$ have the same shape as $\mathbf{z}_H$. Specifically, we parameterize $\mathbf{z}_H$ as a latent tensor of size $C \times F' \times T' = 64 \times 4 \times 16$, where (F', T') are the frequency and time resolutions after the encoder’s strided downsampling of the input spectrogram. The VAE is trained with an ℓ_2 reconstruction loss ($\hat{\mathbf{H}}$ denotes the reconstructed RIR log-mel spectrogram) and a KL regularizer:

$$\mathcal{L}_{\text{vae}} = \|\hat{\mathbf{H}} - \mathbf{H}\|_2^2 + \lambda_1 \text{KL}(q_\phi(\mathbf{z}_H | \mathbf{H}) \| p(\mathbf{z}_H)). \quad (2)$$

3.2. Stage-2: Contrastive Embedding with RIR Alignment

Speech Encoder Architecture. The speech encoder [11] adopts a hybrid CNN–Transformer design, using a strided 2D CNN front-end to encode log-mel spectrograms and a 3-layer Transformer with attention pooling [15] to produce an utterance-level embedding $\mathbf{z}_Y \in \mathbb{R}^D$. We set $D = 4096$ to match the flattened RIR-VAE latent dimensionality for KL-

based alignment. We use cosine similarity for contrastive learning and dispersion; thus we compute an ℓ_2 -normalized embedding $\bar{\mathbf{z}}_Y = \mathbf{z}_Y / \|\mathbf{z}_Y\|_2$. The encoder outputs raw $\mathbf{z}_Y \in \mathbb{R}^D$, which is used for KL alignment and the uncertainty head.

RIR Latent Alignment. This alignment serves as a space regularizer using RIR-only self-supervision (Stage-1) and does not use any downstream task labels or heads. To anchor speech embeddings to the RIR latent space, we align the raw speech embedding $\mathbf{z}_Y$ to the frozen RIR-VAE posterior $q_\phi(\mathbf{z}_H | \mathbf{H}) = \mathcal{N}(\mu_{z_H}, \text{diag}(\sigma_{z_H}^2))$ via a KL term. Following [10, 11], we treat $\mathbf{z}_Y$ as the mean of a Gaussian with fixed identity covariance $\mathbf{I}$, $q_\theta(\mathbf{z}_H | \mathbf{Y}) = \mathcal{N}(\mathbf{z}_Y, \mathbf{I})$. We fix the covariance to $\mathbf{I}$ to avoid introducing an additional uncertainty model in Stage-2; the goal is geometric anchoring rather than probabilistic inference. We minimize

$$\mathcal{L}_{\text{align}} = \text{KL}(q_\theta(\mathbf{z}_H | \mathbf{Y}) \| q_\phi(\mathbf{z}_H | \mathbf{H})). \quad (3)$$

For diagonal Gaussians, the KL has a closed form; we use the mean-over-dim implementation as in [11].

Contrastive Objective. To improve robustness to speech-content variation, we propose to employ a multi-positive contrastive learning approach [16, 17] where positive pairs share the same RIR:

$$\mathcal{L}_{\text{ctr}} = -\frac{1}{N} \sum_{n: |P(n)| > 0} \frac{1}{|P(n)|} \sum_{m \in P(n)} \log \frac{\exp(s_{nm})}{\sum_{k \neq n} \exp(s_{nk})},$$

$$P(n) = \{m \neq n \mid \text{id}_m = \text{id}_n\}, \quad s_{nm} = \frac{\bar{\mathbf{z}}_{Y,n}^\top \bar{\mathbf{z}}_{Y,m}}{\tau}. \quad (4)$$

Here, N is the batch size after flattening, τ is a temperature parameter, and $P(n)$ denotes the set of positives in the batch sharing the same RIR label (id) as anchor n . The loss averages uniformly over all positives for each anchor; anchors without positives are excluded. This objective requires a multi-view batch construction.

The total loss used to train Stage-2 is

$$\mathcal{L}_{\text{encoder}} = \lambda_2 \mathcal{L}_{\text{ctr}} + (1 - \lambda_2) \mathcal{L}_{\text{align}}, \quad \lambda_2 \in [0, 1]. \quad (5)$$

The total loss encourages embeddings to remain discriminative across RIRs while being geometrically consistent with the RIR latent space. This training is supervised by RIR IDs (contrastive + latent alignment) and does not depend on any downstream objective, yielding a task-agnostic representation $\mathbf{z}_Y$.

3.3. Stage-3: Dispersion-Calibrated Uncertainty

In Stage-3, we freeze the encoder and train a lightweight uncertainty head $U = g_\psi(\mathbf{z}_Y)$ to predict the dispersion, without any downstream supervision; U is thus task-agnostic. The head is a two-layer multilayer perceptron with a single hidden layer (size 256) with a Softplus output, ensuring $U \geq 0$.

Embedding Dispersion. A “view” v denotes one corrupted version of the same utterance (noise, frequency mask, or time mask), paired with its uncorrupted clean anchor. For each corrupted sample, dispersion is defined relative to its clean anchor using cosine distance in the ℓ_2 -normalized embedding space:

$$\delta_v = 1 - \cos(\bar{\mathbf{z}}_{Y,v}, \bar{\mathbf{z}}_{Y,\text{clean}}). \quad (6)$$

Training Objective (Rank Loss). In Stage-3, we train the uncertainty head to produce scores that are *order-consistent* with the corruption-induced representation shift, rather than matching its absolute scale. We form pairs (v_1, v_2) of corrupted

samples (optionally restricted to the same corruption type). If $\delta_{v_1} > \delta_{v_2}$, we enforce $U_{v_1} > U_{v_2}$ by a margin-based ranking loss:

$$\mathcal{L}_{\text{rank}} = \mathbb{E}_{(v_1, v_2)} [\max(0, \gamma - (U_{v_1} - U_{v_2}))], \quad (7)$$

where γ is a margin. This objective learns a monotonic mapping from dispersion to uncertainty, avoiding sensitivity to δ 's absolute magnitude. Early stopping is selected by maximizing the Spearman correlation between U and δ on corrupted samples.

4. Experimental Setup

4.1. Datasets

To generate reverberant speech, anechoic utterances were taken from the EARS dataset [18]. A curated collection of measured RIRs was constructed from several publicly available datasets, including the ACE Challenge dataset [2], AIR-IKS [19], ASH-IR [20], Arni [21], BUT ReverbDB [22], EchoThief [23], GTU-RIR [24], MIT IR Survey [25], Multi-Purpose RIR Dataset [26], Multi-Room Transition dataset [27], MeshRIR [28], Motus [29], OpenAIR [30], SoundCam [31], SurrRoom [32], TAU-SRIR [33], and TH Köln dataset [34].

In total, 3000 RIRs ($T_{60} \in [0.046, 1.898]$ s) were collected and randomly partitioned into three mutually exclusive subsets for training, validation, and testing. We split by RIR identity (each measured source–receiver position pair defined one RIR); a room could contribute multiple RIRs at different positions, but no RIR appeared in more than one split. An identical split strategy was applied to all anechoic speech signals. Based on these splits, approximately 89 hours of reverberant speech were generated for training, 11 hours for validation, and 11 hours for testing, using 4-s segments derived from the respective subsets. For Stage-2 training, we used *multi-view* (MV) batches (256 samples): 16 RIRs $\times$ 16 utterances. *Single-view* (SV) batches, as used in [11], contain one utterance per RIR (256 RIRs $\times$ 1; 256 distinct utterances). All RIRs were converted to log-mel spectrograms with 16 mel bands, computed at 16 kHz using a short-time Fourier transform (STFT) with a window length of 64 samples and a hop size of 16 samples (75% overlap). The reverberant speech representations followed the same configuration, except for a hop size of 32 samples (50% overlap).

4.2. Experimental Settings

Speech Corruptions. To probe embedding robustness under varying acoustic conditions, we applied controlled corruptions at waveform and spectrogram levels to the reverberant speech signals. We added waveform-domain pink noise with signal-to-noise ratio (SNR) sampled over speech-active regions: 15 to 25 dB (*mild*), 5 to 15 dB (*medium*), -5 to 5 dB (*severe*). On log-mel spectrograms, we applied SpecAugment-style frequency and time masking [35]. Frequency masking zeroed $k_f = \text{round}(\rho_f F)$ contiguous mel bands, with ρ_f sampled from 5–15% (*mild*), 15–25% (*medium*), or 25–35% (*severe*). Time masking zeroed $k_t = \text{round}(\rho_t T)$ contiguous frames (sequence length preserved), with ρ_t sampled from 0–10% (*mild*), 10–20% (*medium*), or 20–30% (*severe*). The frequency- and time-masking ratios were inspired by the augmentation policies of SpecAugment [35], extended here up to 35% and 30%, respectively. The corruption severity was drawn with probabilities 0.7, 0.25, and 0.05 for *mild*, *medium*, and *severe*, respectively, yielding progressively stronger but mostly mild perturbations for dispersion and uncertainty analysis.

Training hyperparameters. Stage-1 uses $\lambda_1 = 0.05$; Stage-2 uses $\tau = 0.1$ and $\lambda_2 = 1/13$; Stage-3 uses $\gamma = 0.1$.

4.3. Acoustic Parameters

For evaluation, we computed acoustic parameters from the *ground-truth* RIRs and used the values as reference labels. All parameters were computed in $B=7$ octave bands, implemented via a 2nd-order Butterworth filterbank, with center frequencies $f_c^{(b)} \in \{125, 250, 500, 1k, 2k, 4k, 8k\}$ Hz, indexed by $b \in \{1, \dots, B\}$. T_{60} was obtained from a DNN-based decay-slope estimator [27], and C_{50} from the band-limited RIR energy ratio as

$$C_{50}^{(b)} = 10 \log_{10} \left(\frac{\sum_{t=t_{50}-1}^{t_{50}-1} h_{(b)}^2[t]}{\sum_{t=t_{50}}^{L-1} h_{(b)}^2[t]} \right), \quad (8)$$

where $h_{(b)}[t]$ denotes the band-limited RIR in the b -th octave band (length L) at time index t , and t_{50} is the sample index corresponding to 50 ms after the direct-path arrival.

4.4. Performance Metrics

We evaluated (i) the representation quality of Stage-2 embeddings, and (ii) the reliability of the learned uncertainty score U . For representation quality, we report RIR verification [4] (rather than room verification, to assess discrimination across positions within the same room) performance via pair-based retrieval (Average Precision, AP) and reconstruction error of the frozen RIR decoder in the log-mel spectrogram domain (MAE in dB) [10, 11]. For blind acoustic parameter estimation, we report errors for T_{60} and C_{50} (macro-averaged across the 7 standard octave bands).

We also evaluated uncertainty, i.e. how well U reflects corruption-induced representation shift via Spearman correlation between U and embedding dispersion δ . Last, we evaluated *selective prediction*: with samples sorted by U , we show downstream performance with respect to retained coverage.

5. Experimental Results

We report the results for stage-2 representation quality and uncertainty-dispersion consistency as detailed below.

5.1. Stage-2 Representation Quality

Table 1 summarizes Stage-2 representation quality on (a) RIR verification, (b) RIR reconstruction via the frozen RIR-VAE decoder, and (c) blind estimation of T_{60} and C_{50} . For verification, we measure *Average Precision* (AP, area under the precision–recall curve) computed from pairwise cosine similarities, where positives correspond to pairs sharing the same RIR identity and negatives to different RIRs. This metric is designed to reflect the consistency of room embeddings from the same RIR across different speech content. MAE_{rec} denotes the mean absolute error between the decoded and ground-truth log-mel magnitudes, in dB. $\text{MAPE}_{T_{60}}$ denotes the mean absolute percentage error for T_{60} .

We compare against FiNS [6] (time-domain blind RIR estimation from reverberant speech; for reconstruction, its estimated RIR is converted to log-mel magnitudes) and the baseline of Götz *et al.* [11], referred to as MRL (a multi-stage representation learning pipeline with conformal quantile uncertainty). In our setting, MRL-SV is trained with the SV data construction of [11], while MRL-MV keeps the same MRL architecture and objective but replaces SV with our MV data construction; we treat MRL-MV as a data-construction ablation.

Table 1: Stage-2 representation quality and task performance.

Method	AP $\uparrow$	MAE _{rec} (dB) $\downarrow$	MAPE _{T80} (%) $\downarrow$	MAE _{C50} (dB) $\downarrow$
FiNS [6]	0.82	9.59	29.08	2.90
MRL-SV (KL) [11]	0.95	4.76	16.67	1.90
MRL-MV (KL)	0.98	4.04	12.87	1.49
Proposed (Ctr+KL)	0.99	4.06	12.86	1.50

Table 2: Spearman correlation $\rho(U, \delta)$ between uncertainty score U and corruption-induced representation dispersion δ .

Method	Global $\uparrow$	Noise $\uparrow$	F.Mask $\uparrow$	T.Mask $\uparrow$
Severity	-	0.28	0.16	0.17
MRL-MV	0.85	0.59	0.66	0.68
Proposed (U)	0.90	0.83	0.79	0.86

Compared to MRL-SV, our MRL-MV ablation improves verification (AP: 0.98 vs. 0.95), supporting the view that MV speech realizations per RIR improve robustness to speech-content variation under latent alignment. Building on this setting, our proposed model further adds an explicit multi-positive contrastive term, yielding a modest additional gain for verification (AP: 0.99 vs. 0.98). Across reconstruction and blind parameter estimation, results remain similar between MRL-MV and Proposed, suggesting that the contrastive term primarily provides a refinement for verification/retrieval without materially changing performance on the other tasks in Table 1.

5.2. Uncertainty–Dispersion Consistency

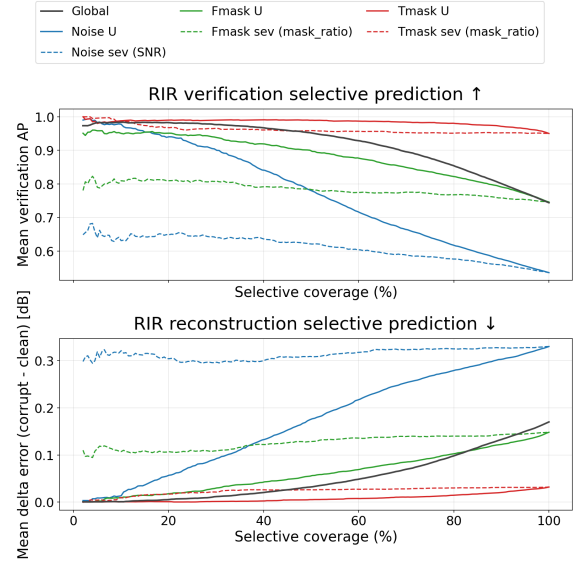
We evaluate whether an uncertainty score is *order-consistent* with corruption-induced representation dispersion, quantified by the clean-anchored dispersion δ (Eq. 6), using Spearman’s rank correlation $\rho(U, \delta)$ (monotonic and scale-free). Importantly, clean anchors are only needed to compute δ (and for training), while inference predicts $U(\mathbf{Y})$ from a single corrupted utterance. We report type-wise correlations for additive noise, frequency masking (F.Mask), and time masking (T.Mask). We include two reference scores: **Severity**, which uses the corruption control parameter (noise: SNR; masking: ratio), and **MRL-MV**, which uses the conformally calibrated inter-quantile range of subband RIR reconstruction errors as uncertainty (retrained under the same corruption protocol). Note that a *global* correlation across corruption types is not reported for Severity, since the control parameters are not comparable across corruption types.

Table 2 shows strong uncertainty–dispersion consistency for our method (global $\rho = 0.90$; robust across corruption types), whereas Severity correlates weakly with dispersion and MRL-MV correlates less strongly than ours, indicating that coarse corruption controls and reconstruction-error uncertainty reflect representation-level instability less reliably under our corruptions.

5.3. Downstream Reliability via Selective Prediction

We assess whether uncertainty U can act as a downstream uncertainty score via *selective prediction*: we rank all corrupted samples by a criterion, retain the top- c fraction with the *lowest* predicted risk, and report performance as a function of coverage $c \in (0, 1]$. U is computed from the corrupted utterance only.

We compare two ranking criteria. **U -sorted** ranks samples in ascending order of U (low $\rightarrow$ high), i.e., we keep samples with the lowest uncertainty first. As an oracle baseline that uses corruption metadata, **severity-sorted** ranks samples by the cor-

Figure 2: Selective prediction. Samples are sorted by uncertainty U (solid) or by corruption severity controls (dashed).

ruption control parameter (noise: descending SNR; masking: ascending mask ratio), i.e., cleaner inputs first. Under both criteria, decreasing coverage removes increasingly corrupted samples; hence, the left side of the curve (small coverage) corresponds to retaining only the cleanest inputs.

Figure 2 shows selective curves for RIR verification AP and reconstruction Δ (corrupt–clean), stratified by corruption type. Across both evaluations, U -sorting yields a steadier improvement as coverage decreases than severity-sorting, indicating that U provides a finer-grained reliability ranking than the coarse corruption controls. This suggests that downstream degradation is not fully explained by the control parameter alone, whereas U better captures the induced representation shift and enables more effective filtering of unreliable samples.

6. Conclusion and Limitations

Building on [11], we improve robustness to speech-content variation in room embeddings learned from reverberant speech and learn a task-agnostic uncertainty score. Our 3-stage pipeline (i) learns a structured RIR latent space using an RIR-VAE, (ii) trains a speech encoder using a multi-view data structure with KL-based alignment to the RIR latent space, adding a multi-positive contrastive term for a modest additional verification gain, and (iii) trains a lightweight uncertainty head via a rank loss to predict corruption-induced embedding dispersion. Experiments under controlled waveform- and spectrogram-level corruptions show that U correlates strongly with representation shift and enables effective selective prediction with single-utterance inference.

While the proposed score supports single-utterance inference and is validated via downstream selective prediction, limitations include that U is a dispersion-calibrated uncertainty score (not a posterior uncertainty) and Stage-3 training relies on clean-anchored dispersion targets (paired clean/corrupted views); our split is by RIR identity (not strictly room-disjoint); and the corruption set (pink noise, SpecAugment masks) does not cover in-the-wild degradations (e.g., interfering speakers, device mismatch, clipping).

7. Generative AI Use Disclosure

Generative AI tools were used only for language editing and polishing (e.g., improving clarity, grammar, and style). All scientific content, including the technical contributions, experimental design, results, and conclusions, was produced and verified by the authors. All authors take full responsibility for the content of this paper and consent to its submission.

8. References

- [1] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.
- [2] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.
- [3] D. Looney and N. D. Gaubitch, "Joint estimation of acoustic parameters from single-microphone speech observations," in *IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, 2020, pp. 431–435.
- [4] Y. Y. Khokhlov, T. Prisyach, A. Mitrofanov, D. Dutov, I. Agafonov, T. Timofeeva, A. Romanenko, and M. Korenevsky, "Classification of room impulse responses and its application for channel verification and diarization," in *Interspeech*, 2024.
- [5] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.
- [6] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 221–225.
- [7] Z. Liao, F. Xiong, J. Luo, M. Cai, E. S. Chng, J. Feng, and X. Zhong, "Blind estimation of room impulse response from monaural reverberant speech with segmental generative neural network," in *Interspeech*, 2023, pp. 2723–2727.
- [8] S. Lee, H.-S. Choi, and K. Lee, "Yet another generative model for room impulse response estimation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023, pp. 1–5.
- [9] Y. Y. Khokhlov, A. Zatzvorskiy, I. Medennikov, I. Sorokin, T. Prisyach, A. Romanenko, A. Mitrofanov, V. Bataev, A. Andrusenko, M. Korenevskaya *et al.*, "R-vectors: New technique for adaptation to room acoustics," in *Interspeech*, 2019, pp. 1243–1247.
- [10] 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 *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 289–293.
- [11] —, "Multi-stage representation learning for blind room-acoustic parameter estimation with uncertainty quantification," *The Journal of the Acoustical Society of America*, vol. 159, no. 1, pp. 247–259, 2026.
- [12] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [13] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1050–1059.
- [14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2014.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [16] 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.
- [17] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 18 661–18 673, 2020.
- [18] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Interspeech*, 2024.
- [19] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *16th International Conference on Digital Signal Processing*, 2009, pp. 1–5.
- [20] S. Pearce, "ASH-IR-Dataset: Acoustic synthetic home impulse response dataset," <https://github.com/ShanonPearce/ASH-IR-Dataset>, 2023, accessed: 2026-02-17.
- [21] K. Pätynen, S. J. Schlecht, and V. Välimäki, "Dataset of impulse responses from variable acoustics room Arni at Aalto Acoustic Labs," 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6985104>
- [22] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.
- [23] C. Warren, "EchoThief Impulse Response Library," <http://www.echothief.com/>, 2024, accessed: 2026-02-17.
- [24] M. Pekmezci, "GTU-RIR: Room impulse response (RIR) data collection and generation project," <https://github.com/mehmetpekmezci/gtu-rir>, 2024, accessed: 2026-02-17.
- [25] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proc. of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.
- [26] L. Friede, C. Tuna, F. Knauff, A. Prinn, K. Ersinadim, and A. Walther, "Multi-Purpose Room Impulse Response Dataset Measured on a 3D Spatial Grid," in *154th Audio Engineering Society Convention*, 2024.
- [27] G. Götz, R. Falcón Pérez, S. J. Schlecht, and V. Pulkki, "Neural network for multi-exponential sound energy decay analysis," *The Journal of the Acoustical Society of America*, vol. 152, no. 2, pp. 942–953, 2022.
- [28] S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, and J. Brunnström, "MeshRIR: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 1–5.
- [29] G. Götz, S. J. Schlecht, and V. Pulkki, "A dataset of higher-order Ambisonic room impulse responses and 3D models measured in a room with varying furniture," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2021, pp. 1–8.
- [30] S. Shelley and D. T. Murphy, "Openair: An interactive auralization web resource and database," in *129th Audio Engineering Society Convention*, 2010, pp. 1270–1278.
- [31] M. Wang, S. Clarke, J.-H. Wang, R. Gao, and J. Wu, "SoundCam: A dataset for finding humans using room acoustics," *Advances in Neural Information Processing Systems*, vol. 36, pp. 52 238–52 264, 2023.
- [32] C. Cieciora, M. Volino, and P. J. B. Jackson, "SurrRoom 1.0 Dataset: Spatial Room Capture with Controlled Acoustic and Optical Measurements," in *154th Audio Engineering Society Convention*. Audio Engineering Society, May 2023.

- [33] A. Politis, S. Adavanne, and T. Virtanen, "TAU spatial room impulse response database (TAU-SRIR DB)," Apr. 2022. [Online]. Available: <https://zenodo.org/records/6408611>
- [34] T. Lübeck, J. M. Arend, and C. Pörschmann, "A high-resolution spatial room impulse response database," *Proc. of the 47th DAGA*, pp. 1–4, 2021.
- [35] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019.