AV-RIR: Audio-Visual Room Impulse Response Estimation

Anton Ratnarajah Sreyan Ghosh Sonal Kumar Purva Chiniya Dinesh Manocha

University of Maryland, College Park, USA

CVPR SEATTLE, WA JUNE 17-21, 2024

Understanding Room Impulse Response

- **What is an RIR?** Room Impulse Response (RIR) measures how sound reflects and decays in a space. Formally, it can be represented as the transfer function between a sound source and a receiver that encapsulates all the direct and reflective paths that sound can travel within any indoor or outdoor environment. We can break down reverberant speech ($\mathcal{S}_R$) into its speech content ($\mathcal{S}_C$) and the corresponding RIR.

$$\mathcal{S}_R = \mathcal{S}_C \circledast \text{RIR},$$

- **What is RIR Estimation?** RIR estimation is the process of determining the RIR from reverberant speech. RIR estimation proves to be important for a variety of applications, including speech recognition, speech enhancement, speech separation, and AR/VR.

Primary Motivation

- Audio-only RIR estimation techniques are capable of estimating early components and are not effective in estimating late components because the early components of the RIR have impulse sparse components, while the late components have a noise-like structure with significantly lower magnitude compared to early components.
- Visual-only RIR estimation demonstrates the feasibility of predicting late components from the RGB image of the environment. However, these approaches are not effective in estimating early components because a single RGB image does not have enough information, such as 3D geometry, information about the material properties of objects in the environment, speaker position, etc.

Main Contributions

- We propose **AV-RIR**, a novel multi-modal multitask learning approach for RIR estimation.
- AV-RIR employs a **neural codec-based multi-modal architecture** that takes as input **audio**, **visual** cues, and a novel **Geo-Mat feature**. We also propose **CRIP** to improve late reverberation effects using retrieval.
- During training, AV-RIR solves an **auxiliary speech dereverberation task** for learning RIR estimation. Through this, AV-RIR essentially learns to separate anechoic speech and RIR. This approach effectively redefines the ultimate learning objective, which is decomposing reverberant speech into its constituent anechoic speech and RIR components.
- We perform extensive experiments to prove the effectiveness of AV-RIR. AV-RIR outperforms prior works by significant margins both quantitatively and qualitatively. We achieve 36% - 63% on RIR estimation on the SoundSpaces dataset, and 56% - 79% people find that AV-RIR is closer to the ground truth in the visual acoustic matching task over our baselines. Additionally, the dereverbed speech predicted by AV-RIR improves performance across various spoken language processing (SLP) tasks. We also perform extensive ablation experiments to demonstrate the critical role of each module within the AV-RIR framework.

AV-RIR Architecture

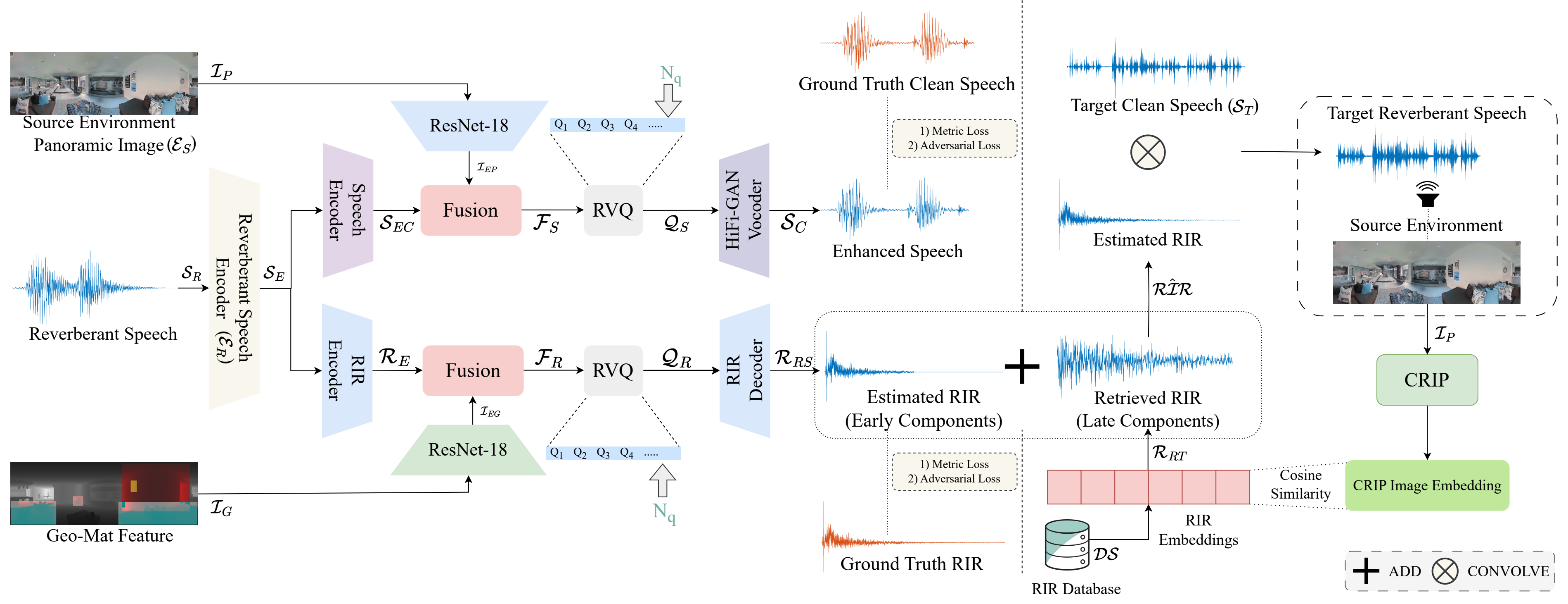


Figure 1. Illustration of AV-RIR.

CRIP & Geo-Mat Feature Map

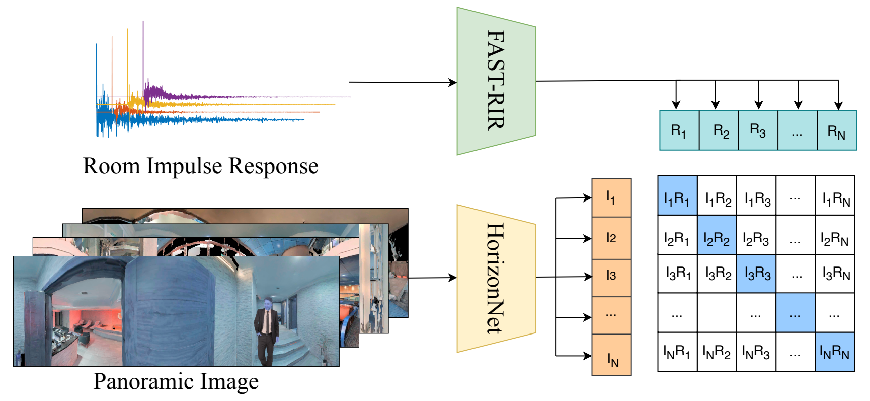

Room Impulse Response
FAST-RIR
R_1 R_2 R_3 ... R_N
Panoramic Image
HorizonNet
I_1 I2 I3 ... I_N
I_1R_1 I_1R_2 I_1R_3 ... I_1R_N
I_2R_1 I_2R_2 I_2R_3 ... I_2R_N
I_3R_1 I_3R_2 I_3R_3 ... I_3R_N
... ...
I_NR_1 I_NR_2 I_NR_3 ... I_NR_N

Figure 2. Illustration of CRIP.

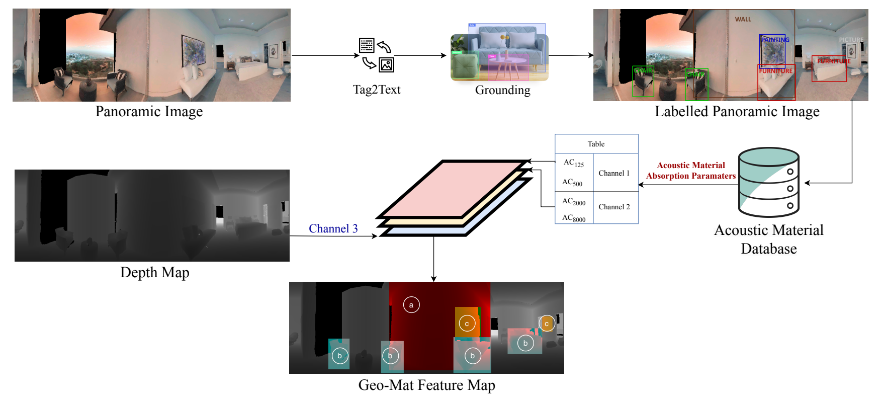
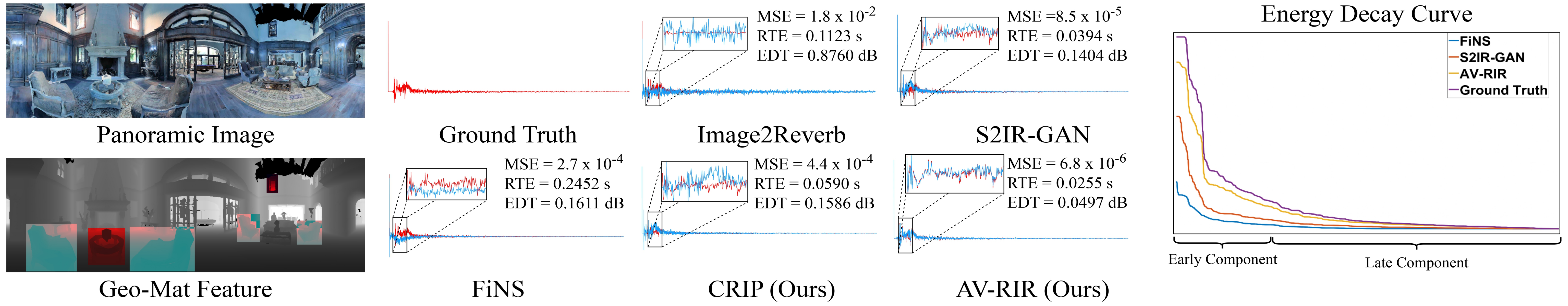


Figure 3. Illustration of Geo-Mat Feature Map

Qualitative Results



Quantitative Results

Method	$\mathbf{T_{60}}$	DRR	EDT	EMSE	LMSE
Image2Reverb	131.7	4.94	382.1	4907	1126
FAST-RIR++	126.4	3.62	334.2	2630	990
FiNS	87.7	3.30	235.7	924	561
S2IR-GAN	63.1	3.04	168.3	730	310
AV-RIR (Audio-Only)	88.8	2.96	122.4	176	51
AV-RIR w/o CRIP	61.7	2.07	79.8	79	42
AV-RIR w/o Geo-Mat	55.7	1.98	74.1	104	6
CRIP-only *(ours)*	118.9	3.14	298.4	212	6
AV-RIR *(ours)*	**40.2**	**1.76**	**62.1**	**82**	**6**

Table 1. Comparison of AV-RIR with prior visual-only and audio-only methods for RIR estimation.

Method	Speech Recognition WER	Speaker Verification EER	RTE (in sec)
Clean (Upper bound)	2.89	1.53	-
Reverberant	8.20	4.51	0.382
MetricGAN+	7.48 (+9%)	4.67 (-4%)	0.187 (+51%)
DEMUCS	7.97 (+3%)	3.82 (+15%)	0.129 (+66%)
HiFi-GAN	9.31 (-14%)	4.32 (+4%)	0.196 (+49%)
WPE	8.43 (-3%)	5.90 (-31%)	0.173 (+55%)
VoiceFixer	5.66 (+31%)	3.76 (+16%)	0.121 (+68%)
SkipConvGAN	7.22 (+12%)	4.86 (-8%)	0.119 (+69%)
Kotha *et al.*	5.32 (+35%)	3.71 (+17%)	0.124 (+68%)
VIDA	4.44 (+46%)	3.97 (+12%)	0.155 (+59%)
AdVerb	**3.54 (+57%)**	3.11 (+31%)	0.101 (+74%)
AV-RIR *(ours)*	4.17 (+49%)	**2.02 (+55%)**	**0.042 (+89%)**

Table 2. Performance comparison of AV-RIR SLU tasks.