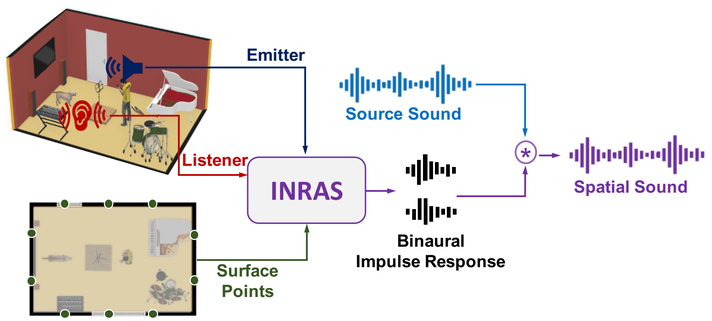
INRAS: Implicit Neural Representation for Audio Scenes

Kun Su[1*], Mingfei Chen[1*], Eli Shlizerman[1,2] *Equal Contribution

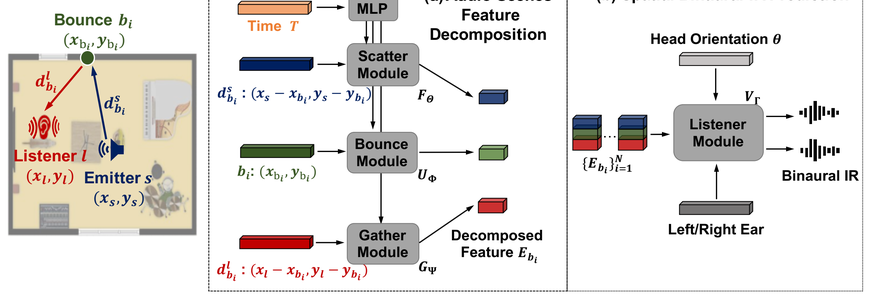
[1] Electrical & Computer Engineering, [2] Applied Mathematics, University of Washington, Seattle, USA

INRAS



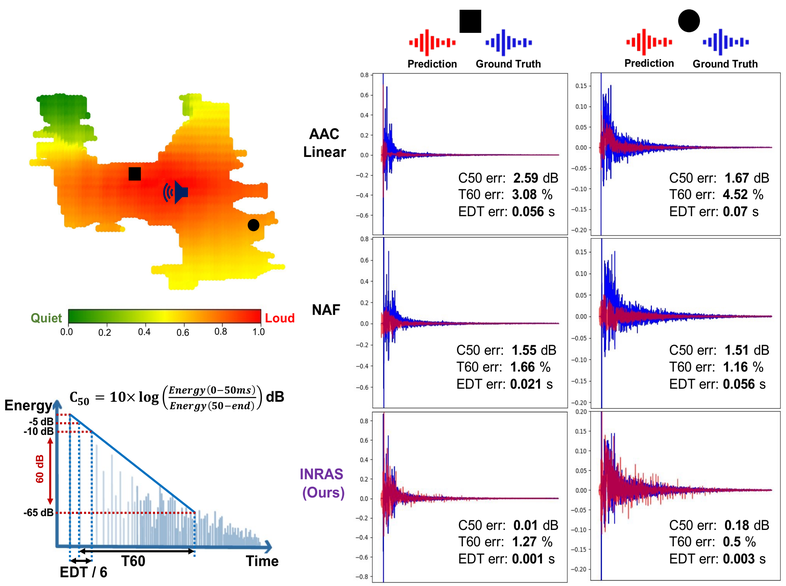
- INRAS learns an implicit neural representation for audio scenes such that given the geometry of a scene, emitter and listener positions, INRAS renders the sound perceived by the listener.
- INRAS outperforms existing approaches on all metrics of audio rendering, including the impulse response quality, inference speed, and storage requirements.

System Overview of INRAS



- **Audio Scenes Feature Decomposition**: Performs audio scenes feature decomposition through three modules: scatter, bounce and gather, corresponding to the three steps of our neural representation. Inputs to the scatter/gather module are the relative distances between the emitter/listener locations and bounce points. The bounce module takes all bounce points to generate scene-dependent features.
- **Spatial Binaural IR Prediction:** In this stage, the decomposed features are stacked and fed to the Listener module, which generates the spatial binaural impulse responses.

Qualitative Results

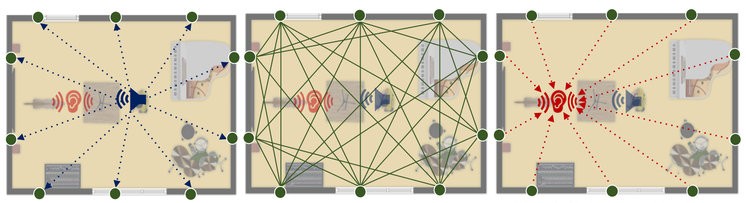


Rendered Impulse Responses Waveform Visualization

- Top left: The speaker indicates the emitter location.
- Right: Rendered waveforms by AAC-Linear, NAF, and INRAS.
- Bottom left: Metrics upon which we evaluate Impulse Responses.

Motivation: Acoustic Radiance Transfer

Motivated by the acoustic radiance transfer, INRAS learns an efficient neural representation for audio scenes according to three steps:



Acoustic radiance transfer steps overview.

- Step 1: Energy scattering from the emitter to all bounce points at the surface.
- Step 2: Acoustic transfer between bounce point pairs.
- Step 3: Energy gathering at the listener from the bounce points.

INRAS disentangles the scene's geometry features to generate independent features for the emitter, the geometry of the scene, and the listener, which leads to the efficient reuse of scene-dependent features for arbitrary emitter-listener positions, and supports effective multi-condition training for multiple scenes by adding only a few trainable parameters.

Evaluations & Results

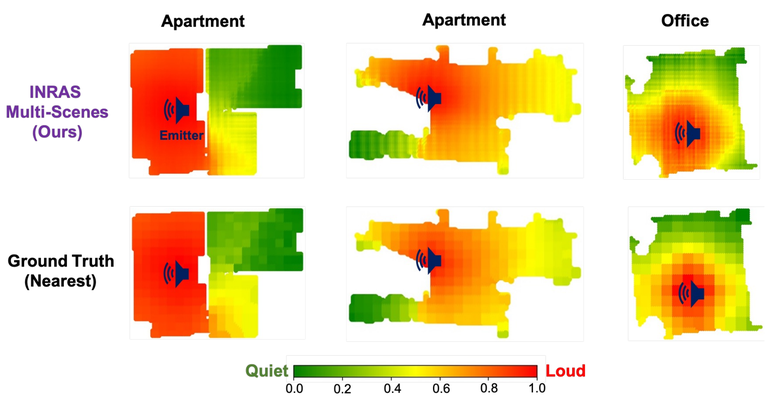
We evaluate our method on the SoundSpaces dataset which includes dense impulse response samples generated by the geometric sound propagation method.

Model\Metric	C50 error (dB) ↓	T60 error (%) ↓	EDT error (sec) ↓	Parameters (Million) ↓	Storage (MB) ↓	Speed (ms) ↓
Opus-nearest	3.58	10.10	0.115	-	181.37	-
Opus-linear	3.13	8.64	0.097	-	181.37	-
AAC-nearest	1.67	9.35	0.059	-	346.74	-
AAC-linear	1.68	7.88	0.057	-	346.74	-
NAF	1.06	3.18	0.031	2.23	8.55	37.86
INRAS (Ours)	**0.6**	**3.14**	**0.019**	**0.67**	**2.56**	**9.47**

Quantitative evaluation of impulse response quality, storage requirements and inference speed. Results are indicated on average of six single-scene models.

Model\Metric	Multi-scenes	SNR (dB) ↑	PSNR (dB) ↑	C50 error (dB) ↓	T60 error (%) ↓	Storage (MB) ↓
Opus-nearest	✗	3.18	13.35	3.6	10.1	544.11
Opus-linear	✗	3.57	13.45	3.23	8.7	544.11
AAC-nearest	✗	6.48	17.84	1.51	9.64	1040.31
AAC-linear	✗	7.52	18.7	1.57	8.05	1040.31
NAF	✗	-1.54	11.25	1.05	**3.01**	25.65
INRAS (Ours)	✓	**8.06**	**18.80**	**0.68**	4.09	**2.99**

Quantitative evaluation of INRAS after multi-condition training on three scene layouts. Results for other methods are computed as an average of the three scenes.



Loudness map visualization

- Top: INRAS multi-scenes rendering of three scenes.
- Bottom: Groundtruth using nearest neighbors.

Visit our poster to see and listen to examples!