ICLR

SonicSim

SonicSim: A customizable simulation platform for speech processing in moving sound source scenarios

Kai Li, Wendi Sang, Chang Zeng, Runxuan Yang, Guo Chen, Xiaolin Hu

Tsinghua University, Beijing, China

National Institute of Informatics, Japan

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Motivation

- **Uncertain Performance in Dynamic Settings**: Existing speech **separation and enhancement** methods excel in static environments, but their performance in dynamic settings is still unknown.
- **Scarcity of Dynamic Source Data**: The **high cost** of recording limits the availability of dynamic source data, hindering the development of speech separation and enhancement techniques in dynamic environments.

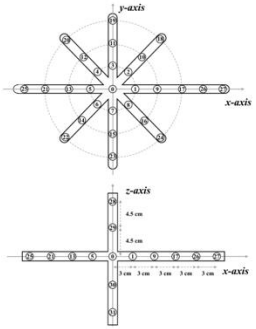


Table 4: Benchmark experiments of speech enhancement.

Baseline	Training Data		Static Speaker						Moving Speaker					
	speech	noise	WB-PESQ	SI-SDR	MOS-SIG	MOS-BAK	MOS-OVR	CER	WB-PESQ	SI-SDR	MOS-SIG	MOS-BAK	MOS-OVR	CER
unprocessed	-		1.15	-9.8	2.00	1.72	1.51	19.9	1.11	-9.1	1.79	1.54	1.36	23.8
FaSNet-TAC [10]	sim	sim	1.38	-3.4	2.67	3.19	2.22	27.1	1.33	-2.5	2.60	3.12	2.14	29.7
	sim	real	1.49	-1.7	**2.83**	3.23	**2.35**	**22.4**	1.42	-1.5	**2.78**	3.20	**2.29**	**25.7**
	real	sim	1.47	0.8	2.67	3.09	2.18	23.7	1.40	0.5	2.58	3.05	2.10	28.2
	real	real	**1.51**	**1.3**	2.80	**3.34**	**2.35**	**22.4**	**1.43**	**1.1**	2.73	**3.28**	2.27	26.3
SpatialNet [28]	sim	sim	1.40	-8.4	**3.09**	2.62	2.28	19.2	1.33	-7.9	**3.06**	2.53	2.23	23.2
	sim	real	1.45	-2.6	2.58	2.35	1.95	23.0	1.38	-2.6	2.54	2.25	1.89	26.5
	real	sim	1.96	3.8	**3.09**	3.06	2.45	17.3	1.80	3.0	3.00	2.99	2.36	**21.2**
	real	real	**2.10**	**6.1**	3.05	**3.51**	**2.62**	**16.0**	**1.90**	**3.8**	2.96	**3.45**	**2.52**	21.5

RealMAN
https://arxiv.org/abs/2406.19959

SonicSim

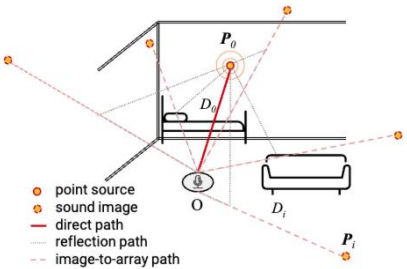
1. 3D Scene Import

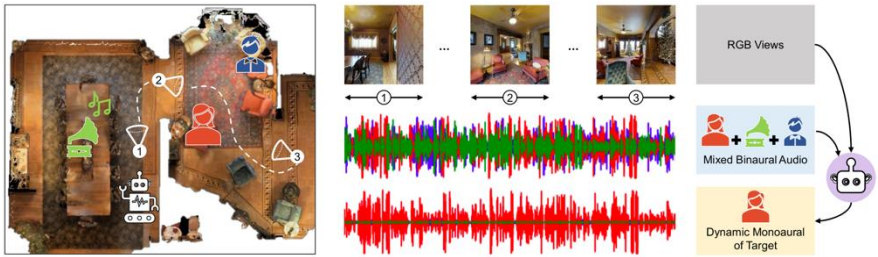
- Imports realistic **3D assets** using Habitat-sim.
- Maintains high fidelity of **geometric data, material properties, and semantic annotations**.
- Simplifies and scales the generation of complex, realistic acoustic environments.

SonicSim

2. Acoustic Environment Simulation

- Simulates sound **reflections** and room acoustics using **path-tracing algorithms**.
- Maps semantic labels to **material properties** (e.g., absorption, scattering).
- Supports **moving sound sources** with dynamic acoustic simulations.

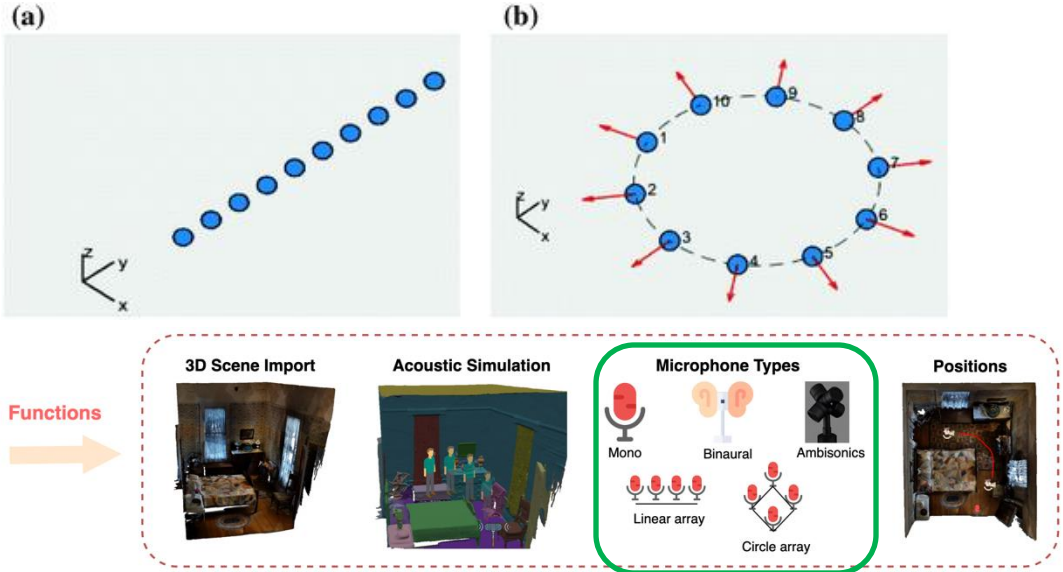




SonicSim

3. Microphone Configurations

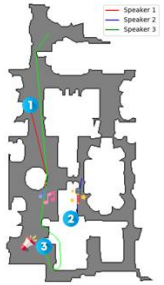
- Supports various audio formats: **mono, binaural, and ambisonics**.
- Allows flexible design of **linear and circular** microphone arrays.
- Provides an API for **custom array layouts** to meet diverse experimental needs.



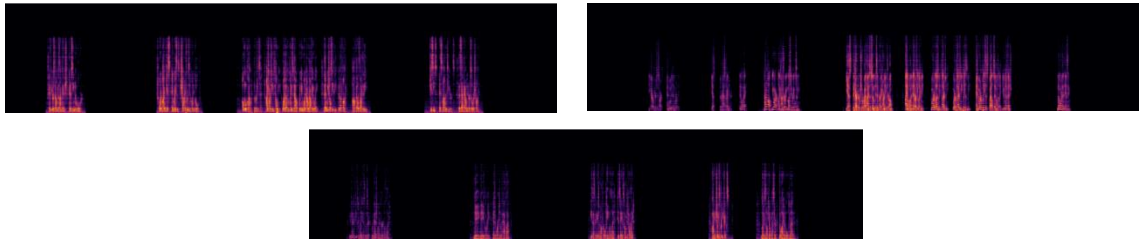
SonicSim

4. Microphone Configurations

- Enables **static or dynamic positioning** of sound sources and microphones.
- Supports **motion trajectories** with real-time acoustic updates.
- Simulates evolving **reverberation, occlusion, and distance effects** dynamically.



Trajectories



Moving audios

SonicSet

1. Multi-Scene

• Composed of **90 diverse scenes** from the Matterport3D dataset, including homes, offices, and churches.

2. Large-Scale

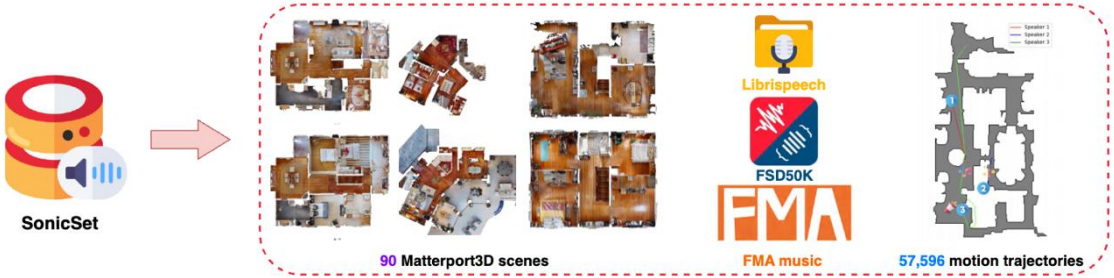
• Integrates **360 hours** of speech audio from LibriSpeech.

• Includes **environmental noise** from FSD50K and musical noise from the FMA dataset.

3. High-Quality

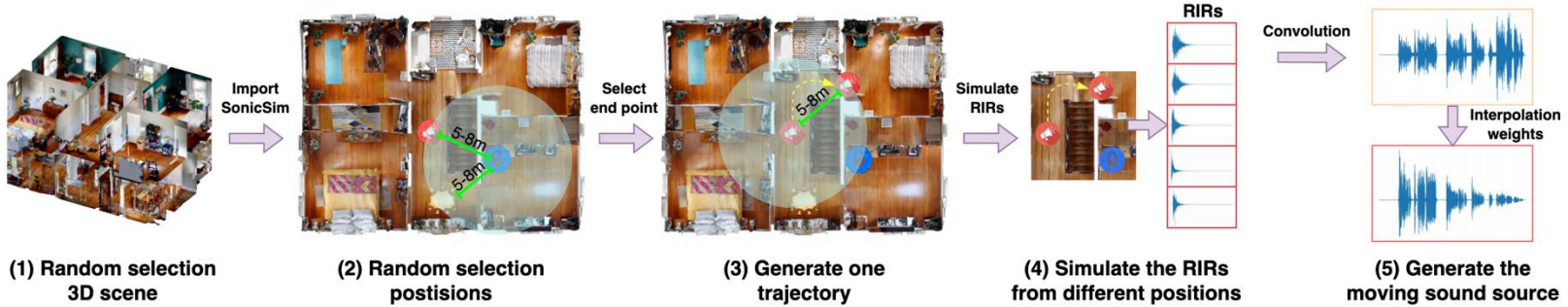
• **Realistic Room Impulse Responses (RIRs)** simulate reflections and diffraction across various materials.

• Results in high-quality, reverberated audio resembling real-world acoustic environments.



SonicSet (Construction method)

1. **Random Selection of 3D Scene**
2. **Random Selection of Positions**
3. **Generate a Trajectory**
4. **Simulate RIRs**
5. **Generate Moving Sound Source**



SonicSet (Compared with other datasets)

Datasets	Geometry	Occlusion	Material	Scalability	Cost	Tools	Src Type	Tasks
WHAMR! 2020	Cuboid	✗	✗	✓	Low	✓	Static	SS/SE
LibriCSS 2020b	Cuboid	✓	✓	✗	High	✗	Static	SS
DNS Challenge 2020	Cuboid	✗	✗	✓	Low	✓	Static	SE
Chime6 2020	Variable	✓	✓	✗	High	✗	Static	SS
LRS2 2023	Variable	✓	✓	✗	High	✗	Static	SS
RealMan 2024	Variable	✓	✓	✗	High	✗	Dynamic	SE
SonicSet (ours)	Variable	✓	✓	✓	Low	✓	Dynamic	SS/SE

Datasets	Speakers	Utterances	Duration (h)	Noise	Reverb	Dynamic
Speech enhancement						
TIMIT (1990)	630	6,300	5	✓	✗	✗
VoiceBank-DEMAND (2016)	110	400	44	✓	✗	✗
DNS Challenge (2020)	~11k	~100k	760	✓	✓	✗
RealMan (2024)	55	-	81	✓	✓	✓
Speech separation						
WSJ0 (2016)	191	28,000	43	✗	✗	✗
Libri2Mix (2020)	1001	56,800	232	✗	✗	✗
LibriCSS (2020b)	40	~1000	10	✓	✓	✗
LRS2-2Mix (2023)	100	48,164	50	✓	✓	✗
Speech separation and enhancement						
WHAM! (2019)	191	28,000	43	✓	✗	✗
WHAMR! (2020)	191	28,000	43	✓	✓	✗
SonicSet (ours)	1001	57,596	960	✓	✓	✓

Real-world Dataset

1. **Audio Selection**

• Randomly selected **30 clean audio samples** from 10 scenes in the SonicSet validation set (5 hours of audio).

2. **Real-World Audio Recording**

• **Playback**: A participant played audio using a 2023 MacBook Pro while moving randomly within the scene.

• **Noise Sources:** Environmental and music noise played from **fixed positions**.

• **Microphone Setup:** Logitech Blue Yeti Nano (omnidirectional, 16 kHz, 32-bit depth) fixed in position.

3. **Data Alignment**

• Clipped audio and noise to match recorded start and stop positions for alignment with original files.

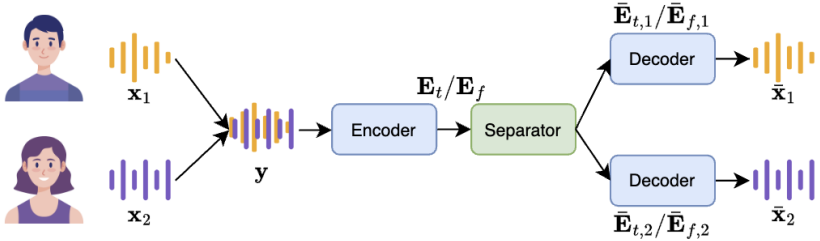
4. **Scene Replication**

• Repeated the same process in **10 similar real-world scenes** using the original audio.

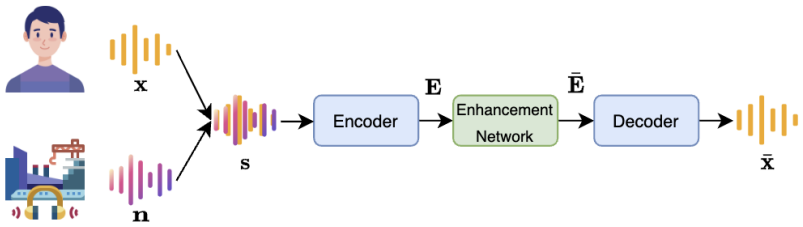
5. **Final Dataset Construction**

• Mixed audio and noise using the **same method** as SonicSet to evaluate model performance.

Speech Separation and Enchancement



(a) Speech separation



(b) Speech enhancement

Speech Separation (test on real-world data)

Method	SI-SNR ↑	SDR ↑	NB-PESQ ↑	WB-PESQ ↑	STOI (%) ↑	WER (%) ↓
Conv-TasNet	2.18/2.45/3.02	3.09/3.24/4.82	1.98/2.03/2.12	1.27/1.31/1.55	59.73/60.33/65.32	98.33/87.04/74.85
DPRNN	2.23/2.81/3.71	2.91/3.54/4.34	1.92/2.05/2.18	1.25/1.32/1.62	60.02/64.76/70.13	91.05/72.63/55.34
DPTNet	4.76/5.53/7.42	5.63/6.68/8.52	2.12/2.32/2.68	1.87/1.91/2.12	71.83/73.42/77.73	51.19/48.18/38.17
SuDoRM-RF	3.44/4.79/5.85	4.22/5.26/6.78	2.08/2.18/2.41	1.58/1.62/1.87	67.77/72.38/73.39	65.22/55.47/48.54
A-FRCNN	3.65/4.87/6.02	4.38/5.67/6.80	2.08/2.21/2.43	1.65/1.68/1.90	69.10/70.23/73.85	68.27/54.32/47.93
SKIM	2.31/2.87/3.33	2.99/3.67/4.13	1.97/2.03/2.07	1.37/1.45/1.63	62.11/64.42/66.67	77.02/70.54/53.84
TDANet	3.90/5.15/6.11	4.71/5.98/7.10	2.15/2.28/2.50	1.72/1.69/1.94	69.95/71.14/74.59	58.40/54.39/43.60
BSRNN	3.68/5.09/6.15	4.46/5.96/6.93	2.10/2.22/2.59	1.79/1.71/2.07	71.26/73.22/76.06	57.63/53.59/48.64
TF-GridNet	6.63/8.27/11.82	7.52/9.10/12.59	2.54/2.71/3.05	2.09/2.28/2.40	79.21/80.34/85.50	34.64/30.21/20.50
Mossformer	5.72/7.94/10.72	6.54/8.78/11.63	2.51/2.60/2.97	2.18/2.23/2.31	75.38/79.32/81.21	47.33/33.84/30.44
Mossformer2	5.87/7.81/10.57	6.66/8.65/11.25	2.56/2.58/2.98	2.23/2.21/2.35	75.50/78.94/81.00	42.94/33.09/29.57

Table 2: Comparative performance evaluation of models trained on different datasets using real-recorded audio with ***environmental noise***. The results are reported separately for "*trained on LRS2-2Mix*", "*trained on Libri2Mix*" and "*trained on SonicSet*", distinguished by a slash. The relative length is indicated below the value by horizontal bars.

Speech Separation (test on SonicSet data)

Method	SI-SNR ↑	SDR ↑	NB-PESQ ↑	WB-PESQ ↑	STOI (%) ↑	WER (%) ↓
Conv-TasNet	4.81 / 4.12	7.13 / 5.38	2.00 / 1.84	1.46 / 1.42	73.14 / 63.21	53.82 / 63.21
DPRNN	4.87 / 4.37	6.65 / 5.73	2.17 / 1.98	1.63 / 1.50	77.63 / 73.73	47.81 / 51.33
DPTNet	11.51 / 11.69	13.00 / 12.80	2.82 / 2.67	2.35 / 2.13	87.62 / 84.23	28.13 / 29.05
SuDoRM-RF	8.01 / 6.84	9.70 / 8.34	2.47 / 2.15	1.98 / 1.66	81.28 / 77.75	35.61 / 41.37
A-FRCNN	9.17 / 7.59	10.63 / 9.32	2.70 / 2.52	2.16 / 2.00	84.82 / 82.14	35.44 / 33.82
SKIM	7.23 / 6.00	8.78 / 7.42	2.34 / 2.23	1.86 / 1.75	79.36 / 77.61	38.92 / 42.82
TDANet	9.27 / 7.00	11.00 / 8.68	2.72 / 2.26	2.22 / 1.71	85.90 / 79.12	30.46 / 37.16
BSRNN	9.10 / 6.96	10.86 / 8.66	2.82 / 2.36	2.26 / 1.76	85.27 / 79.12	29.86 / 41.73
TF-GridNet	15.38 / 14.37	16.81 / 15.69	3.58 / 3.45	3.08 / 2.84	93.32 / 91.80	12.04 / 14.43
Mossformer	14.72 / 11.80	15.97 / 13.17	3.02 / 2.82	2.67 / 2.26	91.13 / 86.15	21.10 / 26.64
Mossformer2	14.84 / 11.12	16.09 / 12.34	3.17 / 2.62	2.83 / 2.09	91.79 / 83.24	19.51 / 32.65

Table 4: Comparison of existing speech separation methods on the SonicSet dataset. The performance of each model is listed separately for results under "*environmental noise*" and "*musical noise*", distinguished by a slash.

Speech Enhancement (test on real-world data)

Method	SDR ↑	WB-PESQ ↑	MOS Sig ↑	MOS Bak ↑	MOS Overall ↑	CER (%) ↓
DCCRN	−1.10/1.87/1.95	1.11/1.24/1.26	2.26/3.25/2.44	2.90/2.12/3.36	1.90/2.19/2.27	50.65/37.56/21.70
Fullband	−1.55/1.18/1.37	1.04/1.07/1.27	2.50/2.84/2.53	2.22/2.61/3.47	2.09/2.19/2.46	51.67/39.71/20.94
FullSubNet	−0.75/1.48/1.92	1.10/1.19/1.30	2.40/2.73/2.69	2.94/2.76/3.48	1.94/2.24/2.46	49.23/32.39/19.15
Fast-FullSubNet	−1.55/1.38/1.37	1.08/1.15/1.31	2.45/3.13/2.67	2.09/2.09/3.48	2.04/1.97/2.59	49.97/40.17/20.08
FullSubNet+	−0.58/1.64/1.32	1.11/1.27/1.28	2.44/2.51/2.59	2.09/2.87/3.52	2.07/2.31/2.46	45.22/23.98/20.48
TaylorSENet	1.06/1.78/2.26	1.21/1.33/1.31	2.44/2.68/2.47	2.09/2.63/2.43	2.10/2.23/2.33	42.55/28.19/20.64
GaGNet	−0.13/1.65/2.10	1.07/1.27/1.30	2.62/2.53/2.46	2.44/3.16/2.41	2.32/2.35/2.40	44.39/34.77/21.09
G2Net	0.01/1.52/2.10	1.10/1.21/1.29	2.76/2.75/2.45	2.21/2.53/2.41	2.07/2.17/2.41	55.12/42.98/21.67
Inter-SubNet	−1.62/1.35/1.61	1.09/1.29/1.34	2.13/2.67/2.70	3.83/2.88/3.47	1.83/2.40/2.51	47.73/22.96/18.73

Table 5: Comparative performance evaluation of models trained on different datasets using the Real-MAN dataset. The results are reported separately for "*trained on VoiceBank-DEMAND*", "*trained on DNS Challenge*" and "*trained on SonicSet*", distinguished by a slash.

Conclusions

- **SonicSim**: A simulation tool for generating **complex acoustic environments** with moving sound sources, integrated with Habitat-sim.
- **SonicSet**: A **large-scale synthetic dataset** designed for speech separation and enhancement tasks.
- Supports multi-scene audio generation.
- Simulates realistic and diverse acoustic environments.
- **Strong Generalization**
- Models pre-trained on SonicSet demonstrate excellent performance on **public benchmarks** and real-world datasets.
- Effectively **bridges the gap** between simulation and real-world scenarios.

Code: https://github.com/JusperLee/SonicSim