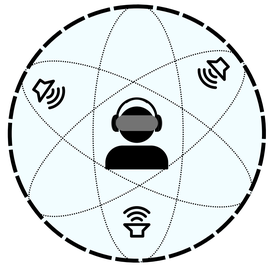
Diff-SAGe: End-to-End Spatial Audio Generation using Diffusion Models

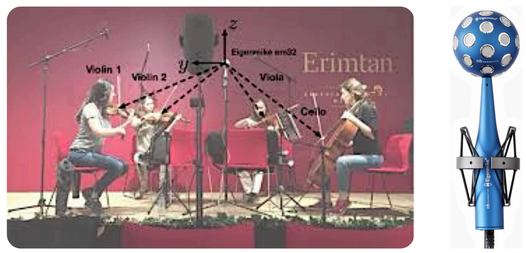
Saksham Singh Kushwaha[1,2,*], Jianbo Ma[2], Mark R. P. Thomas[2], Yapeng Tian[1], Avery Bruni[2]
[1]University of Texas at Dallas, [2]Dolby Laboratories
*(*Work done during internship at Dolby)*

SCAN ME

Introduction

- Spatial audio creates immersive experience.
- Recording is expensive and complex.

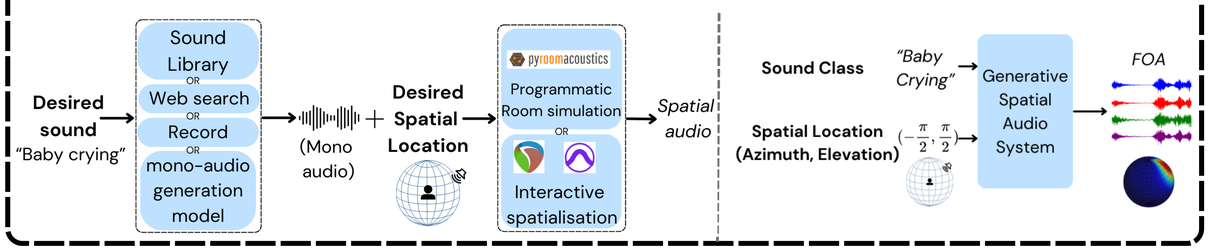




Setup for recording spatial audio

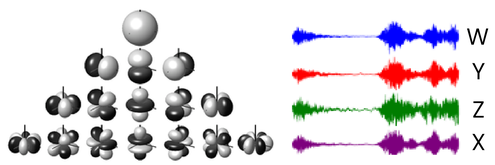
Motivation

- Traditional approaches are not scalable, require expertise etc.
- Our task: end-to-end spatial audio generation
  - Input: Sound class + Spatial location
  - Output: First Order Ambisonics (FOA)

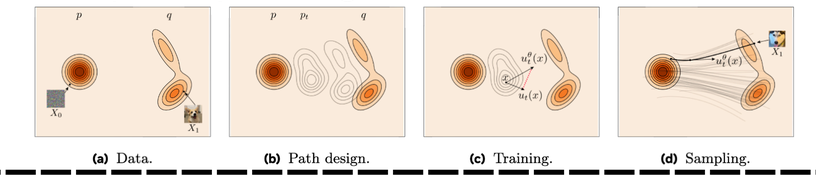


Background

- First Order Ambisonics
  - Mono/Omnidirectional mic: W
  - Directional mics: X, Y, Z



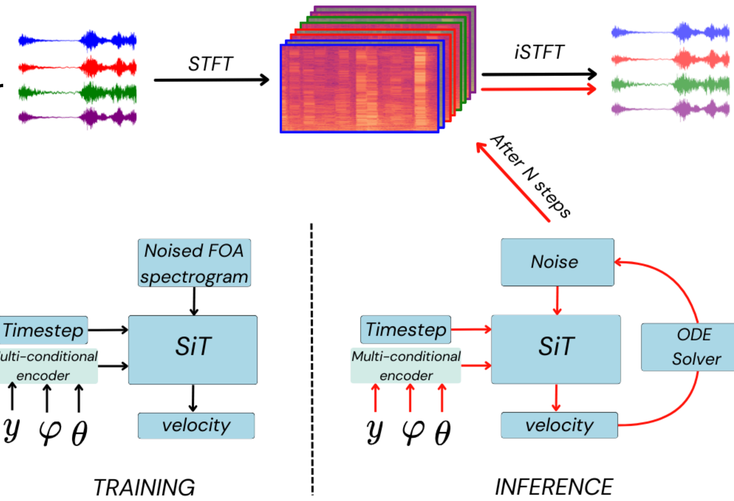
- Flow Matching
  - Learn velocity to map noisy to target data



(a) Data. (b) Path design. (c) Training. (d) Sampling.

Method

- FOA spectrogram
- Multi-conditional encoder
- Flow matching + Diffusion Transformers

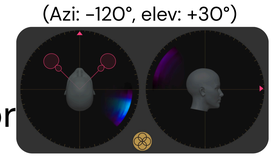


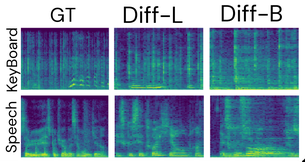
Results

- Datasets:
  - Sound event localisation (SELD)
  - TAU-19, TAU-20
- Baselines:
  - Mono: Real & Generated
  - Simulation: Pyroomacoustics
- Metrics: Condition + Dist. alignment
- DoA Error ⇒ Human annotation error

		TAU-19 Condition		Distribution alignment		
		Acc(%)↑	DoA Error↓	FD↓	FAD↓	KL↓
Ground-Truth	Human	87.19	32.06°	-	-	-
Simulated (Baseline)	Reference audio	62.66	3.33°	10.79	2.47	1.70
	AudioLDM	14.57	**3.22°**	32.62	4.67	2.80
	Tango	35.58	3.92°	23.03	8.04	2.05
(Ours)	Diff-SAGe	**76.52**	22.97°	**3.93**	**0.64**	**1.44**

(Azi: -120°, elev: +30°)





Conclusion

- Introduce a new problem of spatial audio generation.
- Proposed Diffusion transformer (SiT) w/ multi-conditional encoder.
- Surpass simulation based approaches.
- Future directions:
  - Multiple/moving sources; longer duration; 360° video ⇒ FOA

[1] Scheibler et al., "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," ICASSP 2018.
[2] Zotter et al., "Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality." Springer Nature, 2019.
[3] Ma et al., "SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers", ECCV 2024.
[4] Lipman et al., "Flow Matching Guide and Code", arXiv 2024.
[5] Advanne et al. ""A multi-room reverberant dataset for sound event localization and detection", arXiv 2019
[6] DAW and plugin credits: "Ambisonics audio processor Soundfield by Rode" and "Reaper"