Sound Adversarial Audio-Visual Navigation

Yinfeng Yu[1,3], Wenbing Huang[2], Fuchun Sun*[1], Changan Chen[4], Yikai Wang[1,5], Xiaohong Liu[1]

[1]Beijing National Research Center for Information Science and Technology (BNRist),State Key Lab on Intelligent Technology and Systems,Department of Computer Science and Technology, Tsinghua University
[2]Institute for AI Industry Research (AIR), Tsinghua University [3]College of Information Science and Engineering, Xinjiang University [4]UT Austin [5] JD Explore Academy, JD.com

Motivation

Contribution:

- Build an audio simulation platform SoundSpaces[1] to enable audio-visual navigation for two visually realistic 3D environments: Replica[2] and Matterport3D[3].
- Proposed AudioGoal navigation Task:This task requires a robot equipped with a camera and microphones to interact with the environment and navigate to a sounding object.
- SoundSpaces dataset: SoundSpaces is a first-of-its-kind dataset of audio renderings based on geometrical acoustic simulations for two sets of publicly available 3D environments: Replica[2] and Matterport3D[3].

SoundSpaces is focus on audio-visual navigation problem in the acoustically clean or simple environment.

Limitation of SoundSpaces :

- The number of target sound sources is one.
- The position of the target sound source is fixed in an episode of a scene.
- The volume of the target sound source is the same in all episodes of all scenes, and there is no change.

The sound in the setting of SoundSpaces is acoustically clean or simple.

Challenge

However,there are many situations different from the setting of SoundSpaces , which there are some non-target sounding objects in the scene:

For example, a kettle in the kitchen beeps to tell the robotthat the water is boiling, and the robot in the living room needs to navigate to the kitchen and turnoff the stove; while in the living room, two children are playing a game, chuckling loudly fromtime to time.

Challenge 1:

Can an agent still find its way to the destination without being distracted by all non-target sounds around the agent?

non-target sounding objects:

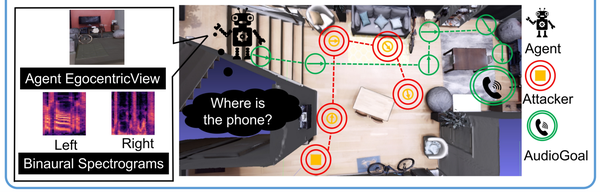
- not deliberately embarrassing the robot: someone walking and chatting past the robot
- deliberately embarrassing the robot: someone blocking the robot forwarding

Challenge 2:

How to model non-target sounding objects in simulator or in reality? There are no such setting existed!

Objecttive

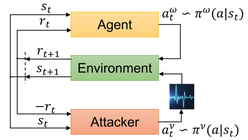
Model non-target sounding objects in simulator.



Mathematical model

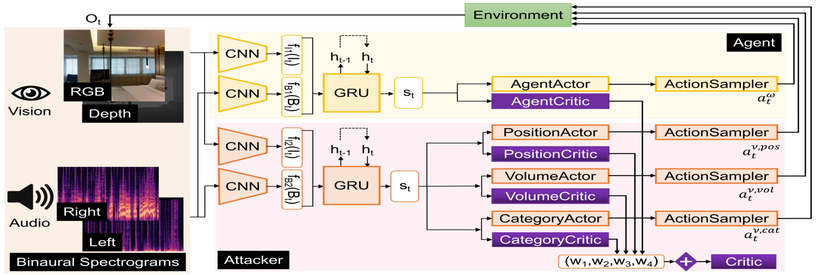
Principle of modeling:

- Worst case strategy: Regard non-target sounding objects as deliberately embarrassing the robot,we called them as sound attacker.
- Simplify:Only consider the simplest situation,one sound attacker.
- Zero sum game:One agent,one sound attacker.



- $G(\pi^{\omega}, \mathrm{r})$ and $G(\pi^{\nu}, \mathrm{r})$ are expected discounted, cumulative rewards of the agent and the attacker respectively.
- $G(\pi^{\omega}, \pi^{\nu}, \mathrm{r}) = [G(\pi^{\omega}, \mathrm{r}), G(\pi^{\nu}, \mathrm{r})]$
- $\pi^{*}=(\pi^{*,\omega}, \pi^{*,\nu}) = \arg\max_{\pi^{\omega}\in[]^{\omega}} \{ \arg\min_{\pi^{\nu}\in[]^{\nu}} \{G(\pi^{\omega}, \pi^{\nu}, \mathrm{r})\}\}$

Neural network model



The agent and the sound attacker first encode observations and learn state representation s_t respectively.

Then, s_t are fed to actor-critic networks, which predicts the next action a_t^{ω} and a_t^{ν} .

Both the agent and the sound attacker receive their rewards from the environment.

The sum of their rewards are zero.

Results

Performance under (SPL (↑)/Rmean (↑)) metrics on Replica.

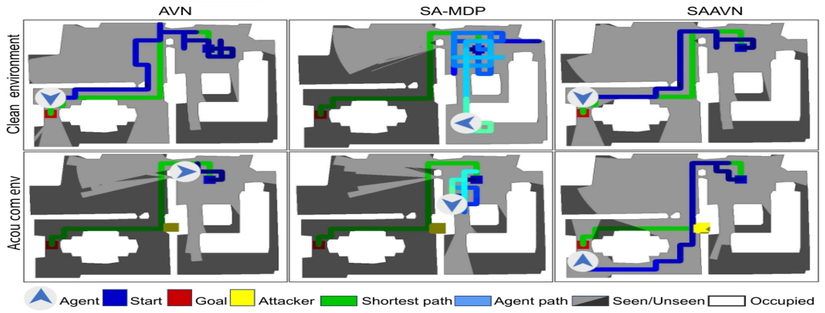
Method	Clean env.	PVC.
Random	0.000/-4.7	0.000/-4.5
AVN	0.721/15.1	0.389/8.0
SA-MDP	0.590/10.2	0.368/7.2
SAAVN	**0.742/16.6**	**0.552/10.6**

Performance under under (SPL (↑)/Rmean (↑)) metrics on Matterport3D.

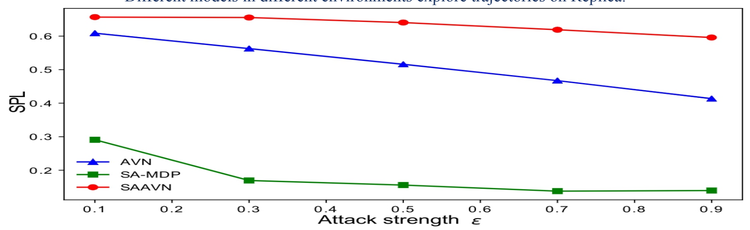
Method	Clean env.	PVC.
Random	0.000/-5.0	0.000/-5.0
AVN	0.539/18.1	0.397/15.3
SAAVN	**0.549/18.7**	**0.478/17.3**

Ablation study
Multi-modal fusion ablation on Replica.

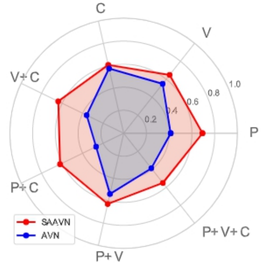
Fusion	SPL (↑)	R_{mean} (↑)
Concatenation	0.552±0.004	10.6±0.1
Element-wise multiply	**0.592±0.005**	**11.8±0.2**



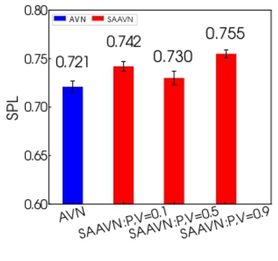
Different models in different environments explore trajectories on Replica.



Navigation capabilities under different sound attack strengths.



Ablation study
Navigation results in acoustically complex environments.



Ablation study
Performance affect by volume.

Conclusion

- This paper proposes a game where an agent competes with a sound attacker in an acoustical intervention environment.
- We have designed various games of different complexity levels by changing the attack policy regarding the position, sound volume, and sound category.
- Interestingly, we find that the policy of an agent trained in acoustically complex environments can still perform promisingly in acoustically simple settings, but not vice versa.
- This observation necessitates our contribution in bridging the gap between audio-visual navigation research and its real-world applications.
- A complete set of ablation studies is also carried out to verify the optimal choice of our model design and training algorithm.

Reference

[1] C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

[2] The Replica Dataset: A Digital Replica of Indoor Spaces, Straub et al., arXiv, 2019

[3] Matterport3D: Learning from RGB-D Data in Indoor Environments, Chang et al., 3DV, 2017

Project & Code

https://yyf17.github.io/SAAVN