Sound Adversarial Audio-Visual Navigation

Yinfeng Yu[1,3], Changan Chen[2], and Fuchun Sun[1⋆]

[1] Department of Computer Science and Technology, Tsinghua University, Beijing, China
yyf17@mails.tsinghua.edu.cn
[2] UT Austin, Texas, America changanvr@gmail.com
[3] College of Information Science and Engineering, Xinjiang University, Urumqi, China

Abstract. Audio-visual navigation task requires an agent to find a sound source in a realistic, unmapped 3D environment by utilizing egocentric audio-visual observations. Existing audio-visual navigation works assume a clean environment that solely contains the target sound, which, however, would not be suitable in most real-world applications due to the unexpected sound noise or intentional interference. In this work, we design an acoustically complex environment in which, besides the target sound, there exists a sound attacker playing a zero-sum game with the agent. More specifically, the attacker can move and change the volume and category of the sound to make the agent suffer from finding the sounding object, while the agent tries to dodge the attack and navigate to the goal under the intervention. Under certain constraints to the attacker, we can improve the robustness of the agent towards unexpected sound attacks in audio-visual navigation. For better convergence, we develop a joint training mechanism by employing the property of a centralized critic with decentralized actors. Experiments on two real-world 3D scan datasets (Replica and Matterport3D) verify the effectiveness and the robustness of the agent trained under our designed environment when transferred to the clean environment or the one containing sound attackers with random policy. Project: https://yyf17.github.io/SAAVN.

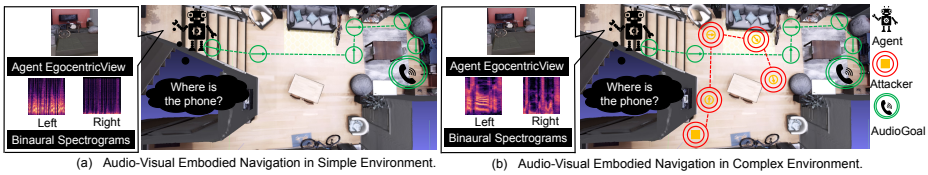
Keywords: Sound adversarial, audio-visual navigation

1 Introduction

Audiovisual embodied navigation, as an important task of embodied vision at present [7, 9, 10], requires agents to find sound source in a real and unmapped 3D environment through egocentric audiovisual observation and exploration [6, 8, 2]. SoundSpaces is the first work to establish an audio-visual embodied navigation simulation platform equipped with the proposed **A**udio-**V**isual embodied **N**avigation (AVN) baseline that resorts to reinforcement learning [4]. The following works for audio-visual embodied navigation are committed to solving long-term exploration [5], sound source is is not periodic and has a variable length [3], and so on.

However, existing audiovisual navigation research results are conducted in the simple setting of a clean environment with only the target sound source. Due to

⋆ Corresponding author: Fuchun Sun.



(a) Audio-Visual Embodied Navigation in Simple Environment. (b) Audio-Visual Embodied Navigation in Complex Environment.

Fig. 1: **Comparison of audio-visual embodied navigation in clean and complex environment.** (a) Audio-visual embodied navigation in an acoustically clean environment: The agent navigates while only hearing the sound emitted by the source object. (b) Audio-visual navigation in an acoustically complex environment: The agent navigates with the audio-visual input from the source object, with the sound attacker making sounds simultaneously.

the existence of moving noise sources such as people talking while walking in the indoor environment, the previous simple settings cannot solve new challenges. The first challenge is how to model non-target moving sounding objects in a simulator or reality? There is no such setting that existed! The second challenge is whether an agent still finds its way to the destination in an acoustically complex environment or not.

We propose first to construct such an acoustically complex environment for the first challenge. In this environment, we add a sound attacker to intervene. The sound attacker can move and change the volume and type of the sound at each time step. In contrast, the agent decides how to move at every time step, tries to dodge the sound attack, and explores for the sound target well under the sound attack, as illustrated in Fig. 1. In reality, most behaviors, such as someone walking and chatting past the robot, are not deliberately embarrassing the robot. How to model this behaviors? Regard non-target sounding objects as deliberately embarrassing the robot under worst case strategy. We called them sound attackers. To simplify, this work only consider the simplest situation, one sound attacker. So the competition between the attacker and the agent can be modeled as a zero-sum two-player game. Our training algorithm is built upon the architecture by [4], with a novel decision-making branch for the attacker. Training two agents separately [13] leads to divergence. Hence we propose a joint Actor-Critic (AC) training framework to solve the second challenge. We define the policies for the attacker based on three types of information: position, sound volume, and sound category. Exciting discoveries from experiments demonstrate that the joint training converges promisingly in contrast to the independent training counterpart. With such a design, we can improve the agent's robustness between the agent and the sound attacker during the game. Our experiments reveal that an agent trained in a worst-case setting can perform promisingly when the environment is acoustically clean or contains a natural sound intervenor using a random policy. On the contrary, the agent trained in a clean environment becomes disabled in an acoustically complex environment.

This work is the first audio-visual navigation method with a sound attacker to the best of our knowledge [14]. To sum up, our contributions are as follows.

– We construct a sound attacker to intervene environment for audio-visual navigation that aims to improve the agent's robustness. In contrast to the environment used by prior experiments [4], our setting better simulates the practical case in which there exist other moving intervenor sounds.
– We develop a joint training paradigm for the agent and the attacker.
– Experiments on two real-world 3D scenes, Replica [12] and Matterport3D [1] validate the effectiveness and robustness of the agent trained under our designed environment when transferred to various cases.

2 Approach

We propose **S**ound **A**dversarial **A**udio-**V**isual **N**avigation (SAAVN), a novel model for the audio-visual embodied navigation task. Our approach is composed of three main modules (Fig. 2). Given visual and audio inputs, our model 1) encodes these cues and make a decision for the motion of the agent, then 2) encodes these cues and decide how to act for the sound attacker to make an acoustically complex environment, and finally 3) make a judgment for the agent and the attacker and to optimization. The agent and the attacker repeat this process until the agent has been reached and executes the Stop action. Our work is based on the SoundSpaces [4] and Habitat [11] and with the publicly available datasets: Replica [12] and Matterport3D [1] and SoundSpaces audio dataset.

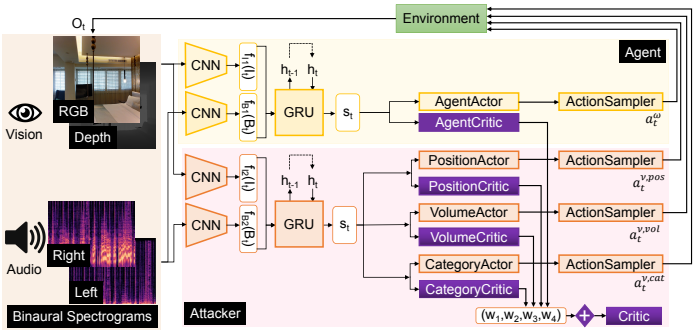


Fig. 2: Sound adversarial audio-visual navigation network. The agent and the sound attacker first encode observations and learn state representation s_t respectively. Then, s_t are fed to actor-critic networks, which predict the next action a_t^{ω} and a_t^{ν}. Both the agent and the sound attacker receive their rewards from the environment.

3 Experiment

Comparison: The effectiveness of our algorithm can be seen through quantitative comparison of performance (see Table 1) and qualitative comparison (see Fig 3).

Table 1: Performance under (SPL (↑)/R_{mean} (↑)) metrics on Replica and Matterport3D . PVC. is a complex Env.

Method	Replica		Matterport3D	
	Clean env.	PVC.	Clean env.	PVC.
Random	0.000/-4.7	0.000/-4.5	0.000/-5.0	0.000/-5.0
AVN	0.721/15.1	0.389/8.0	0.539/18.1	0.397/15.3
SAAVN	**0.742/16.6**	**0.552/10.6**	**0.549/18.7**	**0.478/17.3**

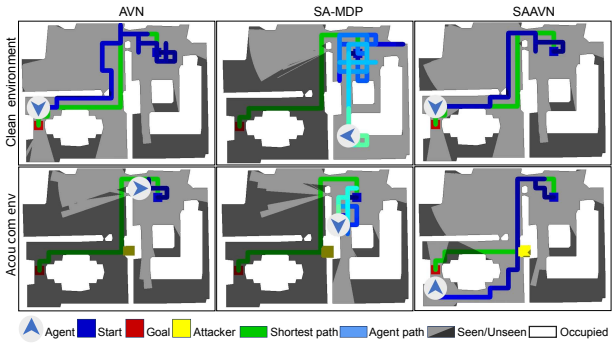


Fig. 3: Different models in different environments explore trajectories. The first row in the figure is a clean environment, and the second line is an acoustically complex environment. Acou com env stands for acoustically complex environment.

Robustness: Fig. 4 shows that our method helps to improve the robust performance.

Ablation study: Fig. 5a demonstrates that SAAVN outperforms AVN in all acoustically complex environments. Fig. 5b reveals that the relationship between the navigation capacity and the volume of the sound attacker is not straightforward and depends on other factors, including the position and sound category.

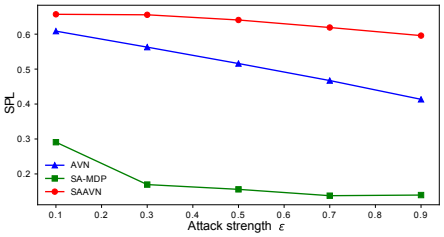
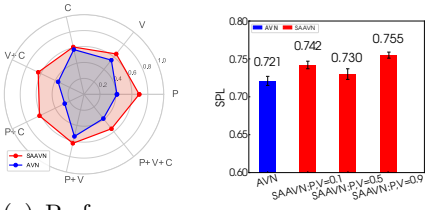


Fig. 4: Performance under different attack strengths.

4 Conclusions

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.



(a) Performance in acoustically complex Env.

(b) Performance affect by volume.

Fig. 5: Ablation study.

References

1. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. International Conference on 3D Vision (3DV) (2017)
2. Chaplot, D.S., Gandhi, D., Gupta, S., Gupta, A., Salakhutdinov, R.: Learning to explore using active neural SLAM. In: ICLR (2020)
3. Chen, C., Al-Halah, Z., Grauman, K.: Semantic audio-visual navigation. In: CVPR (2021)
4. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Soundspaces: Audio-visual navigation in 3d environments. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) ECCV (2020)
5. Chen, C., Majumder, S., Al-Halah, Z., Gao, R., Ramakrishnan, S.K., Grauman, K.: Learning to set waypoints for audio-visual navigation. In: ICLR (2021)
6. Chen, K., de Vicente, J.P., Sepulveda, G., Xia, F., Soto, A., Vázquez, M., Savarese, S.: A behavioral approach to visual navigation with graph localization networks. In: Robotics Science and Systems (2019)
7. Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., Farhadi, A.: Iqa: Visual question answering in interactive environments. In: CVPR (2018)
8. Gupta, S., Davidson, J., Levine, S., Sukthankar, R., Malik, J.: Cognitive mapping and planning for visual navigation. In: CVPR (2017)
9. Lohmann, M., Salvador, J., Kembhavi, A., Mottaghi, R.: Learning about objects by learning to interact with them. In: NeurIPS (2020)
10. Nagarajan, T., Grauman, K.: Learning affordance landscapes for interaction exploration in 3d environments. In: NeurIPS (2020)
11. Savva, M., Malik, J., Parikh, D., Batra, D., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V.: Habitat: A platform for embodied AI research. In: ICCV. pp. 9338–9346 (2019)
12. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
13. Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., Aru, J., Vicente, R.: Multiagent cooperation and competition with deep reinforcement learning. PloS one **12**(4), e0172395 (2017)
14. Yu, Y., Huang, W., Sun, F., Chen, C., Wang, Y., Liu, X.: Sound adversarial audio-visual navigation. In: International Conference on Learning Representations (2022)