Sound Adversarial Audio-Visual Navigation

Yinfeng Yu[1,3],Changan Chen[2], Fuchun Sun[*1]
[1] Beijing National Research Center for Information Science and Technology (BNRist), Department of Computer Science and Technology, Tsinghua University
[2] UT Austin [3] College of Information Science and Engineering, Xinjiang University
yyf17@mails.tsinghua.edu.cn, changanvr@gmail.com

1. Introduction

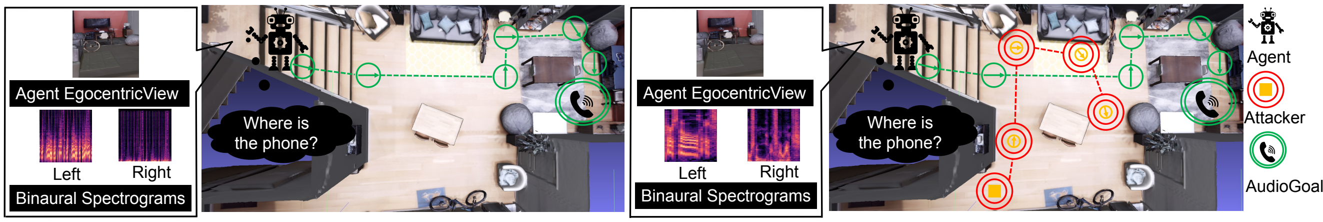
Audiovisual embodied navigation, as an important task of embodied vision at present [12, 14, 15], requires agents to find sound source in a real and unmapped 3D environment through egocentric audiovisual observation and exploration [6, 13, 2]. Inspired by the simultaneous use of eyes and ears in human exploration [20, 8], audio-visual correlation is beneficial to agent learning [11, 9, 7]. A recent work **L**ook, **L**isten, and **A**ct (LLA) has proposed a three-step navigation solution of perception, inference, and decision-making [10]. 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]. In response to the long-term exploration problem that is caused by the large layout of the 3D scene and the long distance to the target place, **A**udio-**V**isual **Wa**ypoint **N**avigation (AV-WaN) proposes an audio-visual navigation algorithm by setting waypoints as sub-goals to facilitate sound source discovering [5]. Besides, **S**emantic **A**udio-**V**isual nav**i**gation (SAVi) develops a navigation algorithm in a scene where the target sound is not periodic and has a variable length; that is, it may stop during the navigation process [3].

However, existing audiovisual navigation research results are conducted in the simple setting of a clean environment with only the target sound source. Due to the existence of moving noise sources such as people talking while walking in the indoor environment, the previous simple settings cannot solve new challenges. For example, the kettle in the kitchen beeps to tell the robot that the water is on, and the robot in the living room needs to navigate to the kitchen to turn off the stove; in the living room, two children are playing games and giggling every now and then. Such examples present a crucial challenge to current technology: can the agent still find its way to the destination without being distracted by all the non-target sounds surrounding the agent? Intuitively, if the agent has not been trained in an acoustically complex environment like the examples listed above, the answer is no. While the answer is no, this capability is something we expect agents to have in real life.

In light of these limitations, we propose first to construct such an acoustically complex environment. In this environment, we add a sound attacker to intervene. This sound attacker can move and change the volume and type of the sound at each time step. In particular, the objective of the sound attacker is to make the agent frustrated by creating a distraction. 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. The competition between the attacker and the agent can be modeled as a zero-sum two-player game. Notably, this is not a fair game and is more biased towards the agent for two reasons. First, the sound attack is just single-modal and will not intervene in any visual information obtained by the agent. Second, as will be specified in our methodology, the sound volume of the attacker is bounded via a relative ratio of the sound target. With such a design, we can improve the agent's robustness between the agent and the sound attacker during the game. On the other hand, our environment is more demanding than reality since there are few attackers in our lives. Instead, most behaviors, such as someone walking and chatting past the robot, are not deliberately embarrassing the robot but just a distraction to the robot, exhibiting weaker intervention strength than our adversarial setting. Even so, 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.

Our training algorithm is built upon the architecture by [4], with a novel decision-making branch for the attacker. Training two agents separately [19] leads to divergence. Hence we propose a joint Actor-Critic (AC) training framework. 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 con-

*Corresponding author: Fuchun Sun.



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

Figure 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.

trast to the independent training counterpart.

This work is the first audio-visual navigation method with a sound attacker to the best of our knowledge [21]. 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 [18] and Matterport3D [1] validate the effectiveness and robustness of the agent trained under our designed environment when transferred to various cases.

2. Overview of Proposed 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.

Environment. Our work is based on the SoundSpaces [4] platform and Habitat simulator [16] and with the publicly available datasets: Replica [18] and Matterport3D [1] and SoundSpaces audio dataset. In SoundSpaces, the sound is created by convolving the selected audio with the corresponding binaural room impulse responses (RIRs) under one of the directions. When a sound attacker emits a chosen sound from its position, the emitted omnidirectional audio is convolved with the corresponding binaural RIR to generate a binaural response from the environment heard by the agent when facing each direction. In this sense, the attacker's sound also considers the reflections on the surface of objects in the environment, making it physically admissible and realistic. The agent's reward is based on how close the robot is away from the goal and whether it succeeds in reaching it. The setting is the same as of the SoundSpaces. The action space of the agent is navigation motions, which is the same as the setting of the SoundSpaces. An environment attacker embodied in the environment must take actions from a hybrid action space $\mathcal{A}^{\nu}$. For brevity, the abbreviation of superscripts position, volume, and category are set to pos, vol, and cat, respectively. The hybrid action space is the Cartesian product of navigation motions space $\mathcal{A}^{\nu,\text{pos}}$, volume of sound space $\mathcal{A}^{\nu,\text{vol}}$ and category of sound space $\mathcal{A}^{\nu,\text{cat}}$: $\mathcal{A}^{\nu} = \mathcal{A}^{\nu,\text{pos}} \times \mathcal{A}^{\nu,\text{vol}} \times \mathcal{A}^{\nu,\text{cat}}$.

Perception, act, and optimization. Our model uses acoustic and visual cues in the 3D environment for efficient navigation. Our model has mainly comprised of three parts: the environment attacker, the agent, and the optimizer (See Fig. 2). At every time step t, the agent and the attacker receives an observation $O_t = (I_t, B_t)$, where I is the egocentric visual observation consisting of an RGB and a depth image; B is the received binaural audio waveform represented as a two-channel spectrogram. Our model encodes each visual and audio observation with a CNN, respectively, where the output of each CNN are visual vector $f_{I1}(I_t)$ and audio vector $f_{B1}(B_t)$. Then, we concatenate the two vectors to obtain observation embedding representation $e^1 = [f_{I1}(I_t), f_{B1}(B_t)]$. We transform observation embedding representation to calculate state representation by a gated recurrent unit (GRU), $s_t^1 = GRU(e_t^1, h_{t-1}^1)$. An actor-critic network uses s_t^1 to predict the action distribution $\pi_\theta^\omega(a_t^\omega | s_t^1, h_{t-1}^1)$ and value of the state $V_\theta^\omega(s_t^1, h_{t-1}^1)$. We also encode visual and audio observation with a CNN for environment attacker, where the output of each CNN are vectors $f_{I2}(I_t)$, $f_{B2}(B_t)$. We then concatenate the two vectors to obtain observation embedding representation $e^2 = [f_{I2}(I_t), f_{B2}(B_t)]$. We also transform observation embedding representation to calculate state

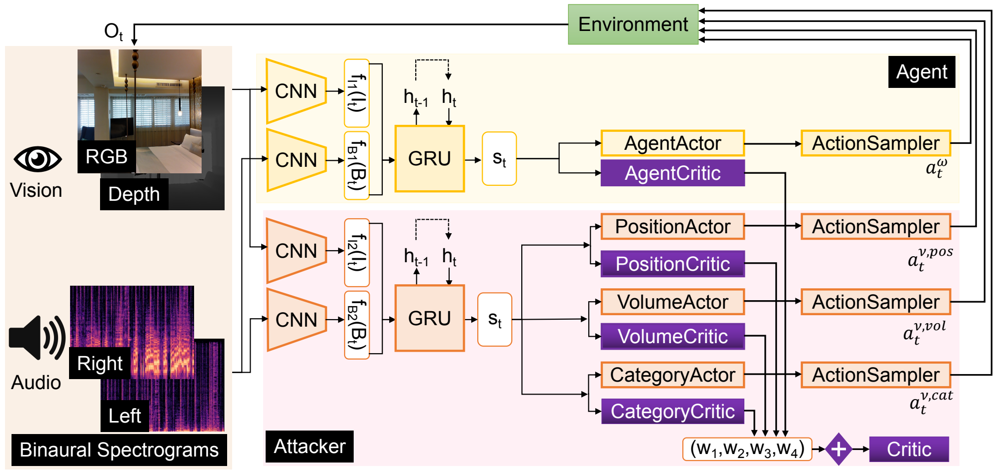


Figure 2. Sound adversarial audio-visual navigation network. The agent and the sound attacker first encode observations and learn state representation s_t respectively. Then, s_t are fed to actor-critic networks, which predict the next action a_t^{ω} and a_t^{ν}. Both the agent and the sound attacker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

representation by a GRU, $s_t^2 = GRU(e_t^2, h_{t-1}^2)$. Three actor-critic networks use s_t^2 to predict the action distribution: $\pi_\theta^{\nu,\text{ pos}}(a_t^{\nu,\text{ pos}}|s_t^2, h_{t-1}^2)$, $\pi_\theta^{\nu,\text{ vol}}(a_t^{\nu,\text{ vol}}|s_t^2, h_{t-1}^2)$, $\pi_\theta^{\nu,\text{ cat}}(a_t^{\nu,\text{ cat}}|s_t^2, h_{t-1}^2)$ and value of the state: $V_\theta^{\nu,\text{ pos}}(s_t^2, h_{t-1}^2)$, $V_\theta^{\nu,\text{ vol}}(s_t^2, h_{t-1}^2)$, $V_\theta^{\nu,\text{ cat}}(s_t^2, h_{t-1}^2)$. All actors and critics are modeled by a single linear layer neural network, respectively. Finally, four action samplers sample the next action a_t^{ω}, $a_t^{\nu,\text{ pos}}$, $a_t^{\nu,\text{ vol}}$, $a_t^{\nu,\text{ cat}}$ from these action distributions generated by AgentActor, PositionActor, VolumeActor and CategoryActor respectively, determining the agent's next motion in the 3D scene. The total critic is a linear sum of PositionCritic, VolumeCritic, and CategoryCritic . The agent and the environment attacker optimize their expected discounted, cumulative rewards $G(\pi^{\omega}, r)$ and $G(\pi^{\nu}, r)$ respectively. The loss of each branch actor-critic network and the total loss of our model as Equation (1).

$$
\begin{aligned}
\mathcal{L}^j &= \sum 0.5 \cdot (\hat{V}_{\theta^j}(s) - V^j(s))^2 \\
&\quad - \sum [\hat{A}^j \log(\pi_{\theta^j}(a \mid s)) + \beta \cdot H(\pi_{\theta^j}(a \mid s))] \\
\mathcal{L}^{\nu} &= 1/3 \cdot (\mathcal{L}^{\nu,\text{cat}} + \mathcal{L}^{\nu,\text{vol}} + \mathcal{L}^{\nu,\text{pos}}) \\
\mathcal{L} &= 1/6 \cdot \mathcal{L}^{\nu,\text{cat}} + 1/6 \cdot \mathcal{L}^{\nu,\text{vol}} + 1/6 \cdot \mathcal{L}^{\nu,\text{pos}} + 1/2 \cdot \mathcal{L}^{\omega}
\end{aligned}
\tag{1}
$$

where $j \in \{(\nu, \text{cat}), (\nu, \text{vol}), (\nu, \text{pos}), (\omega)\}$. $\hat{V}_{\theta^j}(s)$ is estimated state value of the target network for j . $V^j(s) = \max_{a \in \mathbb{A}^j} \mathbb{E}[r_t + \gamma \cdot V^j(s_{t+1}) \mid s_t = s]$. $\hat{A}_t^j = \sum_{i=t}^{T-1} \gamma^{i+2-t} \cdot \delta_i^j$ is the advantage for a given length-T trajectory and $\delta_t^j = r_t + \gamma \cdot V^j(s_{t+1}) - V^j(s_t)$. We optimize the objective follows from *Proximal Policy Optimization* (PPO) [17] .

3. Main Results

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

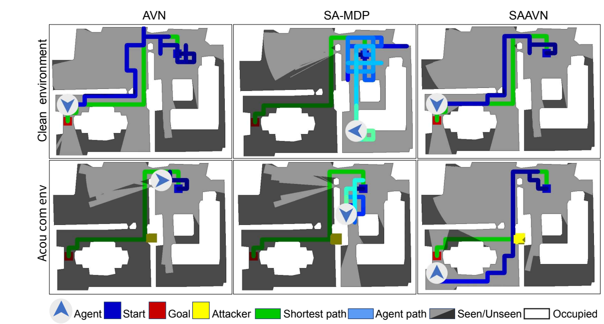


Figure 3. Different models in different environments explore trajectories. The first row in the figure is a clean environment, and the second line is an acoustically complex environment. Acou com env stands for acoustically complex environment.

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**

Robustness: Fig. 4 demonstrates that our method helps to improve the robust performance of the algorithm.

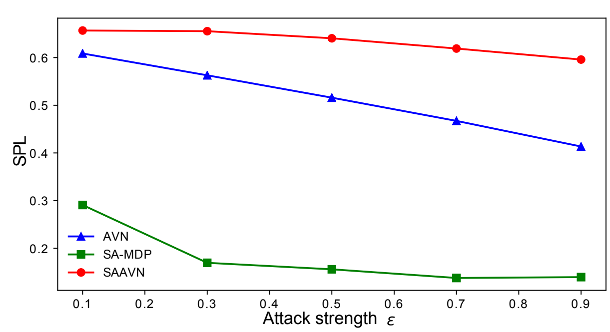


Figure 4. Performance under different attack strengths.

Ablation study: Fig. 5 demonstrates that SAAVN outperforms AVN in all acoustically complex environments. Fig. 6 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. Table 2 show that the new fusion strategy (Element-wise multiply) is better than the original concatenation.

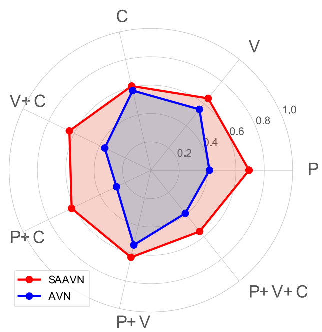


Figure 5. Performance in acoustically complex Env.

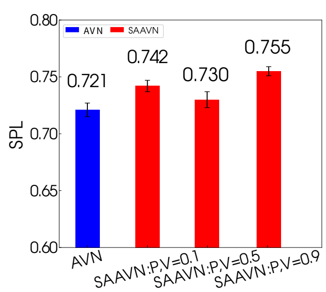


Figure 6. Performance affect by volume.

Table 2. 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**

References

[1] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2

[2] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov. Learning to explore using active neural SLAM. In *ICLR*, 2020. 1

[3] C. Chen, Z. Al-Halah, and K. Grauman. Semantic audio-visual navigation. In *CVPR*, 2021. 1

[4] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman. Soundspaces: Audio-visual navigation in 3d environments. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *ECCV*, 2020. 1, 2

[5] C. Chen, S. Majumder, Z. Al-Halah, R. Gao, S. K. Ramakrishnan, and K. Grauman. Learning to set waypoints for audio-visual navigation. In *ICLR*, 2021. 1

[6] K. Chen, J. P. de Vicente, G. Sepulveda, F. Xia, A. Soto, M. Vázquez, and S. Savarese. A behavioral approach to visual navigation with graph localization networks. In *Robotics Science and Systems*, 2019. 1

[7] V. Dean, S. Tulsiani, and A. Gupta. See, hear, explore: Curiosity via audio-visual association. In *NeurIPS*, 2020. 1

[8] R. Flom and L. Bahrick. The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental psychology*, 43 1:238–52, 2007. 1

[9] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 1

[10] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020. 1

[11] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7062, 2019. 1

[12] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi. Iqa: Visual question answering in interactive environments. In *CVPR*, 2018. 1

[13] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik. Cognitive mapping and planning for visual navigation. In *CVPR*, 2017. 1

[14] M. Lohmann, J. Salvador, A. Kembhavi, and R. Mottaghi. Learning about objects by learning to interact with them. In *NeurIPS*, 2020. 1

[15] T. Nagarajan and K. Grauman. Learning affordance landscapes for interaction exploration in 3d environments. In *NeurIPS*, 2020. 1

[16] M. Savva, J. Malik, D. Parikh, D. Batra, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, and V. Koltun. Habitat: A platform for embodied AI research. In *ICCV*, pages 9338–9346, 2019. 2

[17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[18] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2

[19] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, 12(4):e0172395, 2017. 1

[20] T. Wilcox, R. J. Woods, C. Chapa, and S. McCurry. Multisensory exploration and object individuation in infancy. *Developmental psychology*, 43 2:479–95, 2007. 1

[21] Y. Yu, W. Huang, F. Sun, C. Chen, Y. Wang, and X. Liu. Sound adversarial audio-visual navigation. In *International Conference on Learning Representations*, 2022. 2