SEMANTIC AUDIO-VISUAL NAVIGATION

Changan Chen[1,2] **Ziad Al-Halah**[1] **Kristen Grauman**[1,2]
[1]UT Austin [2]Facebook AI Research

ABSTRACT

Recent work on audio-visual navigation assumes a constantly-sounding target and restricts the role of audio to signaling the target's spatial placement. We introduce *semantic audio-visual navigation*, where objects in the environment make sounds consistent with their semantic meanings (e.g., toilet flushing, door creaking) and acoustic events are sporadic or short in duration. We propose a transformer-based model to tackle this new semantic AudioGoal task, incorporating an inferred goal descriptor that captures both spatial and semantic properties of the target. Our model's persistent multimodal memory enables it to reach the goal even long after the acoustic event stops. In support of the new task, we also expand the SoundSpaces audio simulation platform to provide semantically grounded object sounds for an array of objects in Matterport3D. Our method strongly outperforms existing audio-visual navigation methods by learning to associate semantic, acoustic, and visual cues. Project: `http://vision.cs.utexas.edu/projects/semantic_audio_visual_navigation`. Please see the full paper for more details.

1 INTRODUCTION

An autonomous agent interacts with its environment in a continuous loop of action and perception. The agent needs to reason intelligently about all the senses available to it (sight, hearing, proprioception, touch) to select the proper sequence of actions in order to achieve its task. For example, a service robot of the future may need to locate and fetch an object for a user, go empty the dishwasher when it stops running, or travel to the front hall upon hearing a guest begin speaking there.

Towards such applications, recent progress in visual navigation builds agents that use egocentric vision to travel to a designated point in an unfamiliar environment (Gupta et al., 2017; Savva et al., 2019), search for a specified object (Zhu et al., 2017; Chaplot et al., 2020a), or explore and map a new space (Chen et al., 2019; Chaplot et al., 2020b). Limited new work further explores expanding the sensory suite of the navigating agent to include hearing as well. In particular, methods tackling the AudioGoal challenge use sound to get key directional and distance information about a sounding target to which the agent must navigate (e.g., a ringing phone) (Chen et al., 2020a; Gan et al., 2020; Chen et al., 2020b).

While exciting first steps, existing audio-visual (AV) navigation work has two key limitations. First, prior work assumes the target object constantly makes a steady repeating sound (e.g., alarm chirping, phone ringing). While important, this corresponds to a narrow set of targets; in real-world scenarios, an object may emit a sound only briefly or start and stop dynamically. Second, in current models explored in realistic 3D environment simulators, the sound emitting target has neither a visual embodiment nor any semantic context. Rather, target sound sources are placed arbitrarily in the environment and without relation to the semantics of the scene and objects. As a result, the role of audio is limited to providing a beacon of sound announcing where the object is.

In light of these limitations, we introduce a novel task: *semantic audio-visual navigation*. In this task, the agent must navigate to an object situated contextually in an environment that only makes sound for a certain period of time. Semantic audio-visual navigation widens the set of real-world scenarios to include acoustic events of short temporal duration that are semantically grounded in the environment. It offers new learning challenges. The agent must learn not only how to associate sounds with visual objects, but also how to leverage the semantic priors of objects (along with any acoustic cues) to reason about where the object is likely located in the scene. For example,

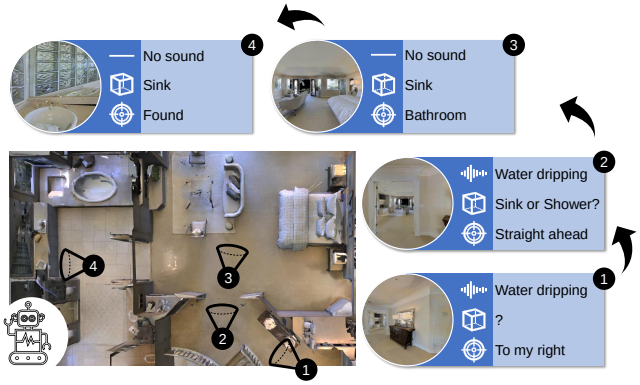

Figure 1: Proposed task: Semantic audio-visual navigation. The agent must navigate to a sounding object in the 3D environment. Since the sound may stop while the agent searches for the object, the agent is incentivized to learn the association between how objects look and sound, and to build contextual models for where different semantic sounds are more likely to occur (e.g., water in the bathroom).

hearing the dishwasher stop running and issue its end of cycle chime should suggest both what visual object to search for as well as the likely paths for finding it, i.e., towards the kitchen rather than the bedroom. Notably, in the proposed task, the agent is not given any external information about the goal (such as a displacement vector or name of the object to search for). Hence the agent must learn to leverage sporadic acoustic cues that may stop at any time as it searches for the source, inferring what visual object likely emitted the sound even after it is silent. See Figure 1.

To tackle semantic AudioGoal, we introduce a deep reinforcement learning model that learns the association between how objects look and how they sound (Fig. 2). Before seeing the target object, the model learns to hypothesize the goal properties (e.g., location and object category) from the received acoustic cues. Coupled with a transformer, it learns to attend to the previous visual and acoustic observations in its memory—conditioned on the predicted goal descriptor—to navigate to the audio source. Furthermore, to support this line of research, we instrument audio-visual simulations for real scanned environments such that semantically relevant sounds are attached to semantically relevant objects.

We evaluate our model on 85 large-scale real-world environments with a variety of semantic objects and their sounds. Our approach outperforms state-of-the-art models in audio-visual navigation with up to 8.9% improvement in SPL (Table 1). Furthermore, our model is robust in handling short acoustic signals emitted by the goal with varying temporal duration, and compared to the competitors, it more often reaches the goal after the acoustic observations end. In addition, our model maintains good performance in the presence of environment noise (distractor sounds) compared to baseline models. Finally, we demonstrate the potential for embodied agents to learn about how objects look and sound through interactions with the 3D environment.

2 Experiments

Baselines. We compare our model to six baselines and existing work: Random, ObjectGoal RL, Chen et al. (2020a), Gan et al. (2020), Chen et al. (2020b) and SMT (Fang et al., 2019) + Audio. All models use the same reward function and inputs. For all methods, there is no actuation noise since audio rendering is only available at grid points (see Chen et al. (2020a) for details).

Metrics. We evaluate the following navigation metrics: 1) success rate: the fraction of successful episodes; 2) success weighted by inverse path length (SPL): the standard metric (Anderson et al., 2018) that weighs successes by their adherence to the shortest path; 3) success weighted by inverse number of actions (SNA) (Chen et al., 2020b): this penalizes collisions and in-place rotations by counting number of actions instead of path lengths; 4) average distance to goal (DTG): the agent's distance to the goal when episodes are finished; 5) success when silent (SWS): the fraction of successful episodes when the agent reaches the goal after the end of the acoustic event.

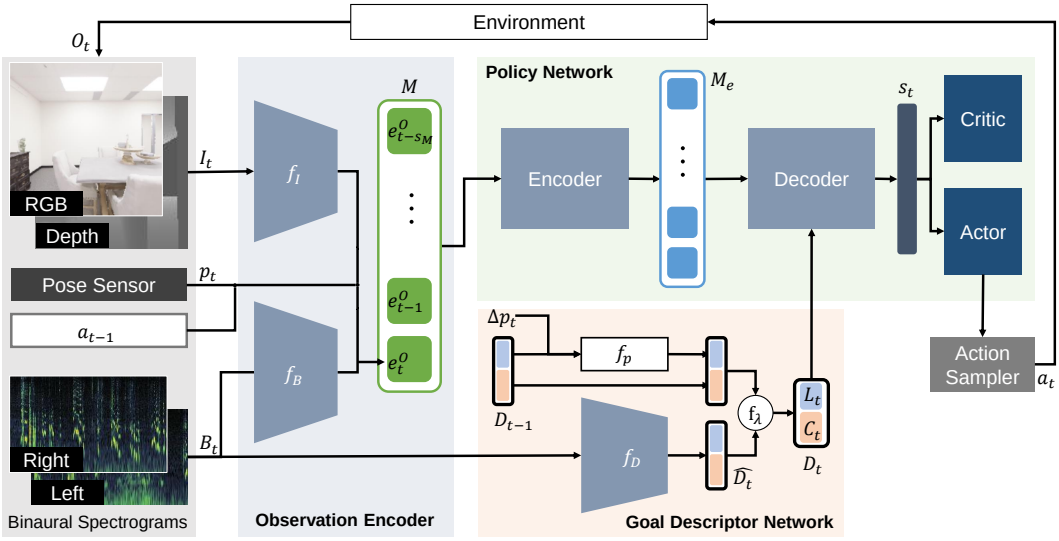

Figure 2: In our model, the agent first encodes input observations and stores their features in memory M. Then it uses the acoustic cues to dynamically infer and update a *goal descriptor* D_t of the target object, which contains both location L_t and object category C_t information about the goal. By conditioning the agent's scene memory on the goal descriptor, the learned state representation s_t preserves information most relevant to the goal. Our transformer-based policy network attends to the encoded observations in M with self-attention to reason about the 3D environment seen so far, and it attends to M_e with D_t to capture possible associations between the hypothesized goal and the visual and acoustic observations to predict the state s_t. Then, s_t is fed to an actor-critic network, which predicts the next action a_t. The agent receives its reward from the environment based on how close to the goal it moves and whether it succeeds in reaching it.

Navigation results. Following the evaluation protocol defined by Chen et al. (2020a), we evaluate all models in two settings: 1) *heard sounds*—train and test on the same sound 2) *unheard sounds*—train and test on disjoint sounds. In both cases, the test environments are always unseen, hence both require generalization. All results are averaged over 1,000 test episodes.

Table 1 shows the results. We refer to our model as **SAVi**: **S**emantic **A**udio-**V**isual Nav**i**gation model. Our approach outperforms all other models by a large margin on all metrics—with 8.9%, 0.3%, 7.2%, 7.2% absolute gains in SPL on *heard sounds* and 3.8%, 4.9%, 4%, 5.3% SPL gains on *unheard sounds* compared to Chen et al. (2020a), Gan et al. (2020), AV-WaN (Chen et al., 2020b), and SMT (Fang et al., 2019), respectively. This shows our model leverages audio-visual cues intelligently and navigates to goals more efficiently. Although Gan et al. (2020) also leverages external supervision for training the location predictor, which leads to good performance on *heard sounds*, this is not enough to solve the task in the more challenging *unheard sounds* setting. Our method has the advantage of allowing the agent to fully leverage the semantic and spatial cues from audio along with its visual perception to locate the sounding objects. According to previously reported results, AV-WaN represents the state-of-the-art for AudioGoal audio-visual navigation. Our SAVi model's gains over AV-WaN show both 1) the distinct new challenges offered by the semantic AudioGoal task, and 2) our model's design effectively handles them.[1]

In addition, our model improves the success-when-silent (SWS) metric by a large margin compared to the closest competitor. This emphasizes the advantage of our goal descriptor module. The explicit and persistent descriptor for the goal in our model helps to maintain the agent's focus on the target even after it stops emitting a sound. Although the SMT+Audio (Fang et al., 2019) model also has access to a large memory pool and can leverage implicit goal information from old observations,

[1]While AV-WaN (Chen et al., 2020b) reports large performance improvements over Chen et al. (2020a) on the standard AudioGoal task, we do not observe similar margins between the two models here. We attribute this to temporal gaps in the memory caused by AV-WaN's waypoint formulation—which are not damaging for constantly sounding targets, but do cause problems for semantic AudioGoal.

	Heard Sounds					*Unheard Sounds*				
	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑	Success ↑	SPL ↑	SNA ↑	DTG ↓	SWS ↑
Random	1.4	3.5	1.2	17.0	1.4	1.4	3.5	1.2	17.0	1.4
ObjectGoal RL	0.9	0.5	0.4	16.7	0.7	0.9	0.5	0.4	16.7	0.7
Chen et al. (2020a)	21.6	15.1	12.1	11.2	10.7	18.0	13.4	12.9	12.9	6.9
Gan et al. (2020)	29.3	23.7	23.0	11.3	14.4	15.9	12.3	11.6	12.7	8.0
AV-WaN	20.9	16.8	16.2	10.3	8.3	17.2	13.2	12.7	11.0	6.9
SMT + Audio	22.0	16.8	16.0	12.4	8.7	16.7	11.9	10.0	12.1	8.5
SAVi (**Ours**)	**33.9**	**24.0**	**18.3**	**8.8**	**21.5**	**24.8**	**17.2**	**13.2**	**9.9**	**14.7**

Table 1: Navigation performance on the SoundSpaces Matterport3D dataset (Chen et al., 2020a). Our SAVi model has higher success rates and follows a shorter trajectory (higher SPL) to the goal compared to the state-of-the-art. Equipped with its explicit goal descriptor and having learned semantically grounded object sounds from training environments, our model is able to reach the goal more efficiently—even after it stops sounding—at a significantly higher rate than the closest competitor (SWS).

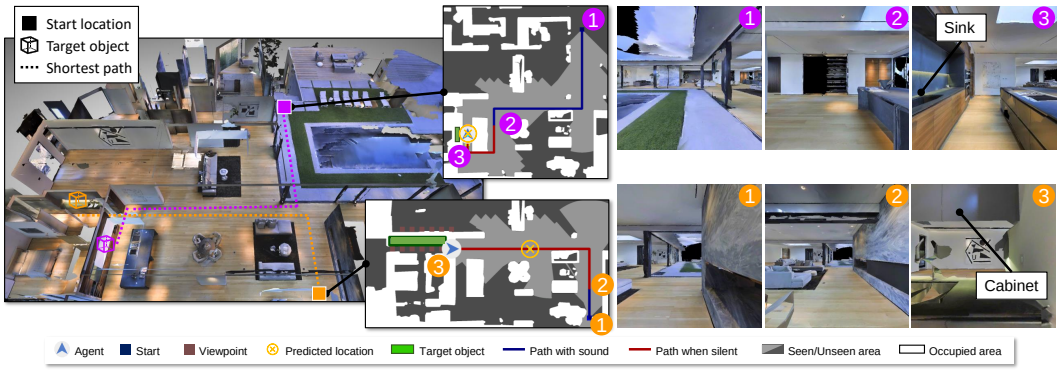

Figure 3: Navigation trajectories for our SAVi model from the test data. In the first episode (top/magenta) the agent hears a water dripping sound and in the second episode (bottom/orange) a sound of opening and closing a door. For each episode, we show three egocentric visual views (right) sampled from the agent's trajectory at the start location ①, when the sound stops ②, and at the end location ③. We see that for the top episode, the acoustic event lasts for two thirds of the trajectory and when the sound stops the agent has an accurate estimate of the object location that helps it find the sounding object (the sink). The second episode (bottom) has a much shorter acoustic event. The agent's estimate of the object location is inaccurate when the sound stops but still helps the agent as a general directional cue. The agent leverages this spatial cue and the semantic cue from its estimate of the object category, a cabinet, to attend to its multimodal memory and find the object in the kitchen and end the episode successfully.

lacking our goal descriptor and the accompanying goal-driven attention, it underperforms our model by a sizeable margin.

As expected, Random does poorly on this task due to the challenging complex environments. Although ObjectGoal RL has the goal's ground truth category label as input, it fails in most cases. This shows that knowing the category label by itself is insufficient to succeed in this task; the agent needs to locate the specific instance of that category, which is difficult without the acoustic cues.

Example navigation trajectories. Figure 3 shows test navigation episodes for our SAVi model. The agent uses its acoustic-visual perception and memory along with the spatial and semantic cues from the acoustic event, whether from a long event (water dripping sound) or a short one (opening and closing a door sound), to successfully find the target objects (the sink and the cabinet). Please see the project website for navigation videos.

REFERENCES

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *NeurIPS*, 2020a. URL http://arxiv.org/abs/2007.00643.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *ICLR*, 2020b.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-visual navigation in 3D environments. In *ECCV*, 2020a.

Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. *arXiv preprint arXiv:2008.09622*, 2020b. URL http://arxiv.org/abs/2008.09622.

Tao Chen, Saurabh Gupta, and Abhinav Gupta. Learning exploration policies for navigation. In *ICLR*, 2019. URL https://openreview.net/pdf?id=SyMWn05F7.

Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *CVPR*, 2019.

Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020.

Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *CVPR*, 2017.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019.

Yuke Zhu, Daniel Gordon, Eric Kolve, Dieter Fox, Li Fei-Fei, Abhinav Gupta, Roozbeh Mottaghi, and Ali Farhadi. Visual Semantic Planning using Deep Successor Representations. In *ICCV*, 2017.