2021 IEEE/CVF International Conference on Computer Vision (ICCV) | 978-1-6654-2812-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCV48922.2021.00034

Move2Hear: Active Audio-Visual Source Separation

Sagnik Majumder[1] Ziad Al-Halah[1] Kristen Grauman[1,2]
[1]The University of Texas at Austin [2]Facebook AI Research
{sagnik, ziad, grauman}@cs.utexas.edu

Abstract

We introduce the active audio-visual source separation problem, where an agent must move intelligently in order to better isolate the sounds coming from an object of interest in its environment. The agent hears multiple audio sources simultaneously (e.g., a person speaking down the hall in a noisy household) and it must use its eyes and ears to automatically separate out the sounds originating from a target object within a limited time budget. Towards this goal, we introduce a reinforcement learning approach that trains movement policies controlling the agent's camera and microphone placement over time, guided by the improvement in predicted audio separation quality. We demonstrate our approach in scenarios motivated by both augmented reality (system is already co-located with the target object) and mobile robotics (agent begins arbitrarily far from the target object). Using state-of-the-art realistic audio-visual simulations in 3D environments, we demonstrate our model's ability to find minimal movement sequences with maximal payoff for audio source separation. Project: http://vision.cs.utexas.edu/projects/move2hear.

1. Introduction

Audio-visual events play an important role in our daily lives. However, in real-world scenarios, physical factors can either restrict or facilitate our ability to perceive them. For example, a father working upstairs might move to the top of the staircase to better hear what his child is calling out to him from below; a traveler in a busy airport may shift closer to the gate agent to catch the flight delay announcements amidst the din, without moving too far in order to keep her suitcase in sight; a friend across the table in a noisy restaurant may tilt her head to hear the dinner conversation more clearly, or scooch her chair to better catch music from the band onstage.

Such examples show how *controlled sensor movements* can be critical for audio-visual understanding. In terms of audio sensing, a person's nearness and orientation relative to a sound source affects the clarity with which it is heard, especially when there are other competing sounds in the environment. In terms of visual sensing, one must see obstacles to circumvent them, spot desired and distracting sound sources, use visual context to hypothesize an out-of-view sound source's location, and actively look for "sweet spots" in the visible 3D scene that may permit better listening.

Figure 1: Active audio-visual source separation. Given multiple mixed audio sources S_i in a 3D environment, the agent is tasked to separate a target source (shown in green) by intelligently moving around using cues from its egocentric audio-visual input to improve the quality of the predicted target audio signal. See text.

In this work, we explore how autonomous multi-modal systems might learn to exhibit such intelligent behaviors. In particular, we introduce the task of *active audio-visual source separation*: given a stream of egocentric audio-visual observations, an agent must decide how to move in order to recover the sounds being emitted by some target object, and it must do so within bounded time. See Figure 1. Unlike traditional audio-visual source separation, where the goal is to isolate sounds in passive, pre-recorded video [23, 27, 1, 20, 43, 28, 17, 29], the proposed task calls for actively controlling the camera and microphones' positions over time. Unlike recent embodied AI work on audio-visual navigation, where the goal is to travel to the location of a sounding object [15, 25, 14, 16], the proposed task calls for returning a separated soundtrack for an object of interest in limited time, without necessarily traveling all the way to it.

We consider two variants of the new task. In the first, the system begins exactly at the location of the desired sounding object and must fine-tune its positioning to hear better; this variant is motivated by augmented reality (AR) applications where the object of interest is known and visible (e.g., the person seated across from someone wearing an assistive

978-1-6654-2812-5/21/$31.00 ©2021 IEEE
DOI 10.1109/ICCV48922.2021.00034

audio-visual AR device) yet local movements of the device sensors are still beneficial to improve the audio separation. In the second variant, the system begins at an arbitrary position away from the object of interest; this variant is motivated by mobile robotics applications, where an agent detects a sound from afar (e.g., the child calling from downstairs) but it is entangled with distractors and requires larger movements within the environment to hear correctly. We refer to these scenarios as *near-target* and *far-target*, respectively.

Towards addressing the active audio-visual source separation problem, we propose Move2Hear, a reinforcement learning (RL) framework where an agent learns a policy for how to move to hear better. The agent receives an egocentric audio-visual observation sequence (RGB and binaural audio) along with the target category of interest[1], and at each time step decides its next motion (a rotation or translation of the camera and microphones). During training, as it aggregates these observations over time with an explicit memory module and recurrent network, the agent is rewarded for improving its estimate of the target object's latent (monaural) sound. In particular, the reward promotes movements that better resolve the target sound from the other distractor sounds in the environment. Our approach handles both near- and far-target scenarios.

Importantly, optimal positioning for audio source separation is *not* the same as navigation to the source, both because the agent faces a time budget—it may be impossible to reach the target in that time—and also because the geometry of the 3D environment and relative positions of distractor sounds make certain positions relative to the target more amenable to separation. For example, in Fig. 1 the agent is tasked with separating the audio from object S_3. Here, going directly to S_3 is not ideal, as that position will have high interference from the other audio sources in the scene, S_1 and S_2. By moving around the kitchen bar, the agent manages to dampen the signal from S_1 significantly due to the intermediate obstacles (walls) and, simultaneously, emphasize the signal from S_3 compared to S_2, thus leading to better separation.

We test our approach with realistic audio-visual SoundSpaces [15] simulations on the Habitat platform [54] comprising 47 real-world Matterport3D environment scans [10], along with an array of sounds from diverse human speakers, music, and other background distractors. Our model successfully learns how to move to hear its target more clearly in unseen scenes, surpassing baselines that systematically survey their surroundings, smartly or randomly explore, or navigate directly to the source. We explore the synergy of both vision and audio for solving this task.

Our main contributions are 1) we define the active audio-visual separation task, a new direction for embodied AI research; 2) we present the first approach to begin tackling this task, namely a new RL-based framework that integrates sound separation and visual navigation motion policies; and 3) we thoroughly experiment with a variety of sounds, visual environments, and use cases. While just the first step in this area, we believe our work lays groundwork to explore new problems for multi-modal agents that move to hear better.

[1] e.g., a human speaker, a musical instrument, or some other sound type

2. Related Work

Passive Audio(-Visual) Source Separation. Passive (non-embodied) separation of audio sources using solely audio inputs has been extensively studied in signal processing. While sometimes only single-channel monaural audio is assumed [59, 61, 69, 34], multi-channel audio captured with multiple microphones [39, 80, 19]—including binaural audio [18, 70, 76]—facilitates separation by making the spatial cues explicit. Using vision together with sound improves separation. Audio-visual (AV) separation methods leverage mutual information [33, 21], subspace analysis [60, 46], matrix factorization [44, 56, 27], correlated onsets [6, 35], and deep learning to separate speech [1, 23, 20, 43, 2, 17, 29], music [28, 24, 74, 77], and other objects [27]. While some methods extract the audio tracks for all classes present in the mixture [57, 66], others isolate one specific target [42, 67, 30, 31, 81].

Whereas prior work assumes a pre-recorded video as input, our work addresses a new *embodied perception* version of the audio separation task, in which an agent can see, hear, and move in a 3D environment to actively hear a source better. To our knowledge, our work is the first to consider how intelligent movement influences a multi-modal mobile agent's ability to separate sound sources. In addition, whereas existing video methods use dynamic object motion to tease out audio-visual associations (especially for speech [23, 1, 20, 43, 17, 29]), our setting demands using visual cues in the surrounding 3D environment to move to "sweet spots" for listening to a source amidst competing background sounds. Finally, our task requires recovering the target object's latent monaural audio as output—stripping away the effects of the agent's relative position, environment geometry, and scene materials. This aspect of the task is by definition absent for AV separation in passive video.

Visual and Audio-Visual Navigation. While mobile robots traditionally navigate by a mixture of explicit mapping and planning [65, 22], recent work explores learning navigation policies from egocentric image observations (e.g., [32, 53, 36]). Facilitated by fast rendering platforms [54] and realistic 3D visual assets [10, 73, 62], researchers develop reinforcement learning architectures to tackle diverse visual navigation tasks [32, 36, 71, 78, 75, 7, 72, 37, 13, 11, 79, 53, 12]. Going beyond purely visual agents, recent work explores joint audio-visual sensing for embodied AI [25, 15, 26, 47, 14, 16]. In the *audio-visual navigation* task, an agent enters an unmapped environment

and must travel to a sounding target object (e.g., go to the ringing phone) [25, 15, 14, 16]. Related efforts explore audio-visual spatial sensing to infer environment geometry [26] or floorplans [47], or attempt audio-only navigation to multiple fixed-position sources in a gridworld while accounting for distractor sounds [49].

Our *separation* goal is distinct from *navigation*: our agent succeeds if it accurately separates the true target sound, not if it simply travels to where the sound or target object is. As we will show, the two tasks yield agents with differing behavior. In fact, our model remains relevant even in the near-target scenario, where (unlike AV navigation) the position of the target is already known. Compared to any of the above, our key novel insight is that audio-visual cues can tell an agent how to move to separate multiple active audio sources.

Source Localization in Robotics Robotic systems use microphone arrays to perform sound source localization, often via signal processing techniques on the audio stream alone [41, 50]. To focus on a sound, like a human speaker, the microphone can be actively steered towards the localized source (e.g., [4, 40, 9]). Using both visual and audio cues to localize people and detect when they are speaking [3, 68, 5] is an important precursor to human-robot systems that follow conversations. The proposed task also requires actively attending to audio events, but in our case there are multiple competing sound sources and they may be initially far from the agent. Our technical contribution is complementary to existing methods: our approach learns to map audio-visual egocentric observations directly to long-term sequential actions. Learning behaviors from data, as opposed to fixing heuristics, offers potential advantages for generalization.

3. Active Audio-Visual Source Separation

We propose a novel task: active audio-visual source separation (AAViSS). In this task, an autonomous agent simultaneously hears multiple audio sources of different types (e.g., speech, music, background noise) that are at various locations in a 3D environment. The agent's goal is to separate a *target source* (e.g., a specified human speaker or instrument) from the heard audio mixture by intelligently moving in the environment. This *active listening* task requires the agent to leverage both acoustic and visual cues. While the acoustic signal carries rich information about the types of audio sources and their relative distances and orientations from the agent, the visual signal is crucial both to see obstacles affecting navigation and identify useful locations from which to sample acoustic information in the visible 3D scene.

Task Definition. In each episode of agent experience, multiple audio sources are randomly initialized in the 3D environment. The map of the environment is unknown to the agent as are the locations of the audio sources. At each step, the agent hears a mixed binaural audio signal that is a function of the source types (e.g., human voice, music, etc.), their displacement from the agent, and their sound reflections resulting from the major geometric surfaces and materials in the 3D scene. One of the audio sources is the target, i.e., the source the agent wants to hear, as relevant to its overarching application setting. The agent is tasked with predicting the target's *monaural* audio signal as clearly as possible—that is, the true latent target sound itself, separated from the other sources and independent of the spatial effects of where it is emitted. The agent must intelligently move and sample visual-acoustic cues from its environment to best predict the target signal by the end of a fixed time budget.

Note that it is significant that we define the correct output to be the *monaural* target sound. Were the objective instead to output the *binaural* sound of the target at the agent's current position, trivial but non-useful solutions would exist (e.g., moving to a position where the target is inaudible, and hence its binaural waveforms are approximately 0).

As discussed above, we consider two variants of this task depending on the agent's starting position relative to the target audio source. In the *near-target* variant, the agent starts at the target and needs to conduct a sequence of fine-grained motions to extract the best target audio; in the *far-target* variant, the agent starts in a random far position and must first navigate to the vicinity of the target before commencing movements for better separation.

Episode Specification. Formally, an episode is defined by a tuple $(E, p_0, S_1, S_2, \ldots S_k, G^y)$ where E is a 3D scene, $p_0 = (l_0, r_0)$ is the initial agent pose defined by its location l and rotation r, $S_i = (S_i^w, S_i^l, S_i^y)$ is an audio source defined by its periodic monaural waveform S_i^w, its location S_i^l, and type S_i^y. There are k audio sources in the scene, each from a different type[2] ($S_1^y \neq S_2^y \neq \ldots \neq S_k^y$), and G^y is the target audio goal type such that $G \in \{S_i\}$. At each step, the agent hears a mixture of all audio sources, and the goal is to predict G^w by the end of the episode, given the target goal label G^y. The episode length is $\mathcal{T}$ steps, meaning the agent has bounded time to provide its output.

Action Space. The agent's action space $\mathcal{A}$ consists of $MoveForward$, $TurnLeft$, and $TurnRight$. At each step, the agent samples an action $a_t \in \mathcal{A}$ to move on a *navigability graph* of the environment; that graph is unknown to the agent. While turning actions are always valid, $MoveForward$ is allowed only if there is an edge connecting the current node and the next one, and the agent is facing the destination node. Non-navigable connections exist due to walls and obstacles, e.g., a sofa blocking the path.

3D Environment and Audio-Visual Simulator. Consistent with substantial embodied AI research in the computer

[2]Note that distinct human voices count as distinct types.

vision community (e.g., [15, 48, 32, 53]), and in order to provide reproducible results, we develop our approach using state-of-the-art visually and acoustically realistic 3D simulators. We use the SoundSpaces [15] audio simulations, built on top of the AI-Habitat simulator [54] and the Matterport3D scenes [10]. The Matterport3D scenes are real-world homes and other indoor environments with both 3D meshes and image scans. SoundSpaces provides room impulse responses (RIR) at a spatial resolution of 1 meter for the Matterport3D scenes. These state-of-the-art RIRs capture how sound from each source propagates and interacts with the surrounding geometry and materials, modeling all of the major real-world features of the RIR: direct sounds, early specular/diffuse reflections, reverberations, binaural spatialization, and frequency dependent effects from materials and air absorption. Our experiments further push the realism by considering noisy sensors (Sec. 5). See the Supp. video to gauge realism.

We place k monaural audio sources in the 3D environment. Since current simulators do not support dynamic object rendering (e.g., people talking, instruments being played), these sources are represented as point objects [15].

At each time step, we simulate the binaural mixture of the k sounds coming from their respective locations in the scene, as received by the agent at its current position. Specifically, the sources' waveforms S_i^w are convolved with RIRs corresponding to the scene E and the agent pose and the source location pairs (p, S_i^l). Subsequently, the output of the convolved RIRs is mixed together to generate dynamic binaural mixtures as the agent moves around:

$$B_t^{w,mix} = \frac{1}{k}\sum_{i=1}^{k} B_t^{w,i}, \tag{1}$$

where $B_t^{w,i}$ is the binaural waveform of the sound source i at time t, and $B_t^{w,mix}$ is the binaural waveform of the mixed audio. Note that the ground truth waveforms per source are known only by the simulator; during inference, the agent observes only the mixed binaural sound, which changes with t as a function of the agent's movements.

4. Move2Hear Approach

We pose the AAViSS task as a reinforcement learning problem, where the agent learns a policy to sequentially decide how to move given its stream of egocentric audio-visual observations. Our model has two main components (see Fig. 2): 1) the target audio separator network and 2) the active audio-visual (AV) controller. The separator network has two functions: it separates the target audio signal from the heard mixture at each step, and it informs the controller about its current estimate to improve the separation. The AV controller learns a policy guided by the separation quality, such that it moves the agent in the 3D environment to improve the predicted target audio signal. These two components learn from each other during training to help the

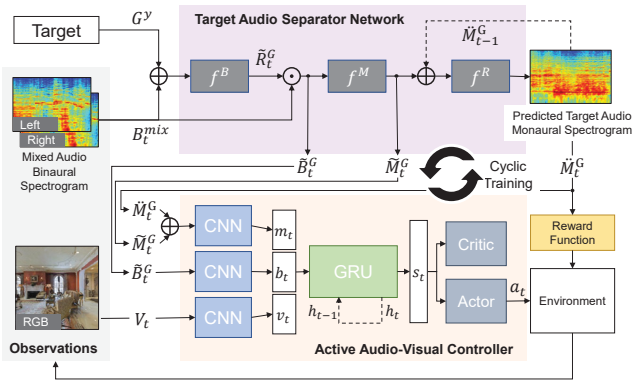


Figure 2: Our model for active audio-visual source separation has two main components: 1) an audio separator network (top) and 2) an active audio-visual controller (bottom). At each step, our model receives mixed audio from multiple sources in the 3D environment along with egocentric RGB views. The model actively moves in the environment to improve its separation of a target audio source.

agent build an implicit understanding of how the separation quality changes given the current surrounding 3D structure (furniture, walls, rooms, etc.), the various audio sources, and their (unobserved) locations relative to the agent and the target. Hence, they enable the agent to learn a useful active movement policy to improve audio separation quality.

4.1. Target Audio Separator Network

At each step t, the audio network f^A receives the mixed binaural sound B_t^{mix} coming from all audio sources in the scene and the target audio type G^y, and it predicts the monaural target audio $\ddot{M}^G$, i.e., $f^A(B_t^{mix}, G^y) = \ddot{M}^G$ (Fig. 2 top). We use the short-time Fourier transform (STFT) to represent both the monaural M and binaural B audio spectrograms. M and B are matrices, with $B \in \mathbb{R}_+^{2\times F\times N}$ and $M \in \mathbb{R}_+^{F\times N}$ where F is the number of frequency bins, N is the time window, and B has two channels (left and right).

The audio network f^A predicts $\ddot{M}^G$ in three steps and using three modules, such that: $f^A = f^B \circ f^M \circ f^R$. First, given the target category G^y, the binaural target audio separator f^B separates the target's binaural signal $\tilde{B}_t^G$ from the input mixture B_t^{mix}. Second, the monaural audio predictor f^M takes the previous binaural output $\tilde{B}_t^G$ and predicts the monaural target audio $\tilde{M}_t^G$ (i.e., independent of the room acoustics and spatial effects). Finally, given the monaural estimates from all previous steps and the current one $\tilde{M}_t^G$, the acoustic memory refiner f^R continuously enhances the target monaural audio prediction $\ddot{M}_t^G$. Next we describe the architecture of these three modules in detail.

Binaural Audio Separator. For the binaural extractor f^B, we use a multi-layer convolutional network in a U-Net like architecture [51] with ReLU activations in the last layer. Specifically, we concatenate the target label G^y with the

mixed binaural audio B^{mix} along the channel dimensions, and pass this input to the U-Net to predict a real-valued ratio mask $\tilde{R}^T$ [28, 74, 77] for the target binaural spectrogram:

$$\tilde{R}^G_t = f^B(B^{mix}_t \oplus G^y), \tag{2}$$

where $\oplus$ denotes channel-wise concatenation. Then, the predicted spectrogram for the target binaural $\tilde{B}^G_t$ is obtained by soft-masking the mixed binaural spectrogram with $\tilde{R}^G_t$:

$$\tilde{B}^G_t = \tilde{R}^G_t \odot B^{mix}_t, \tag{3}$$

where $\odot$ denotes the element-wise product of two tensors.

Monaural Audio Predictor. Similarly, we use another U-Net for f^M to predict the target monaural audio at the current step given the prediction from f^B, i.e.:

$$\tilde{M}^G_t = f^M(\tilde{B}^G_t). \tag{4}$$

$\tilde{M}^G_t$ serves as an initial estimate of the target monaural that our model predicts based only on the mixed binaural audio heard at the current step t. The factorization of monaural prediction into two steps f^B and f^M allows our model to first focus on the audio source separation by extracting the target audio in the same domain as the input (binaural) using f^B then to learn how to remove spatial effects from the binaural to get the monaural signal using f^M. Additionally, f^B provides the policy with spatial cues about the target to help the agent anchor its actions in relation to the target location (see Sec. 4.2). We refine this prediction using an acoustic memory, as we will describe next.

Acoustic Memory Refiner. The acoustic memory refiner f^R is a CNN that receives the current monaural separation $\tilde{M}^G_t$ from f^M and its own previous prediction $\ddot{M}^G_{t-1}$ as inputs, and predicts the refined monaural audio $\ddot{M}^G_t$:

$$\ddot{M}^G_t = f^R(\tilde{M}^G_t \oplus \ddot{M}^G_{t-1}). \tag{5}$$

See Fig. 2 top right. The acoustic memory plays an important role in stabilizing the monaural predictions and helping the agent learn a useful policy, and it also provides robustness to microphone noise. By taking into consideration the previous estimate $\ddot{M}^G_{t-1}$ and its relation to $\tilde{M}^G_t$, the model can learn when to update the monaural estimate of the target and by how much. This encourages non-myopic behavior: it allows the agent to explore its vicinity with less pressure to reduce the quality of the predictions, in case there is a need to traverse intermediary low quality spots in the environment. Consequently, when the navigation policy is trained with a reward that is a function of the improvement in f^R's prediction quality (details below), the agent learns to visit spaces in the scene that will improve the separation quality over time.

All three modules f^B, f^M, and f^R are trained in a supervised manner using the ground truth of the target binaural and monaural signal, as detailed in Sec. 4.3.

4.2. Active Audio-Visual Controller

The second component of our approach is an AV controller that guides the agent in the 3D environment to improve its audio predictions (Fig. 2 bottom). The controller leverages both visual and acoustic cues to predict a sequence of actions a_t that will improve the output of f^A, the separator network defined above. It has two main modules: 1) an observation encoder and 2) a policy network.

Observation Space and Encoding. At every time step t, the AV controller receives the egocentric RGB image V_t, the current binaural separation $\tilde{B}^G_t$ from f^B, and the channel-wise concatenation of the target monaural predictions from f^M and f^R, that is, $\bar{M}^G_t = \tilde{M}^G_t \oplus \ddot{M}^G_t$.

The audio and visual inputs carry complementary cues required for efficient navigation to improve the separation quality. $\tilde{B}^G_t$ conveys spatial cues about the target (its general direction and distance relative to the agent) which helps the agent to anchor its actions in relation to the target location. Importantly, the better the separation quality of $\tilde{B}^G_t$, the more apparent this directional signal is (compared to B^{mix}_t, the mixed audio directly observed by the agent). $\bar{M}^G_t$ is particularly useful in letting the policy learn the associations between the model's current position and the quality of the prediction in that position (captured by $\tilde{M}^G$), the overall quality so far (captured by $\ddot{M}^G$), and whether this position led to a change in the estimate (captured by $\bar{M}^G$).

The visual signal V_t provides the policy with cues about the geometric layout of the 3D scene so that the agent can avoid colliding with obstacles. Further, the visual input coupled with the audio allow the agent to capture relations between the 3D scene and the expected separation quality at different locations.

We encode the three types of input using separate CNN encoders: $v_t = E^V(V_t)$, $b_t = E^B(\tilde{B}^T_t)$ and $m_t = E^M(\bar{M}^G_t)$. We concatenate the three feature outputs to obtain the current audio-visual encoding $o_t = [v_t, b_t, m_t]$.

Policy Network. The policy network is made up of a gated recurrent network (GRU) that receives the current audio-visual encoding o_t and the accumulated history of states h_{t-1} to update its history to h_t and outputs the current state representation s_t. This is followed by an actor-critic module that takes s_t and h_{t-1} as inputs to predict the policy distribution $\pi_\theta(a_t|s_t, h_{t-1})$ and the value of the state $\mathcal{V}_\theta(s_t, h_{t-1})$, where θ are the policy parameters. The agent samples an action $a_t \in \mathcal{A}$ according to π_θ to interact with its environment.

Near- and Far-Target Policies. For the near-target task, we learn a *quality policy* $\pi^{\mathcal{Q}}$ that is driven by the improvement in the target audio prediction (reward defined below). For the far-target variant, we learn a composite policy made up of $\pi^{\mathcal{Q}}$ and an *audio-visual navigation policy* $\pi^{\mathcal{N}}$ trained to get closer to the target audio.

The composite policy uses a time-based strategy to switch control between the two policies. First, the navigation policy brings the agent closer to the target audio in a budget of $\mathcal{T}^{\mathcal{N}}$ steps, then the agent switches control to the quality policy to focus on improving the target audio separation. We found alternative blending strategies, e.g., switching based on predicted distance to the target, inferior in practice. The audio network f^A is active throughout the episode in both the near- and far-target tasks.

4.3. Training

Training the Target Audio Separator Network. The separator network has two outputs: the binaural $\tilde{B}$ and the monaural audio predictions $\tilde{M}$ and $\ddot{M}$. We train it using the respective ground truth spectrograms of the target audio which are provided by the simulator:

$$\mathcal{L}^B = ||\tilde{B}_t^G - B_t^G||_1, \tag{6}$$

where B_t^G is the ground truth binaural spectrogram of the target at step t. Similarly, for the monaural predictions:

$$\mathcal{L}^M = ||\tilde{M}_t^G - M^G||_1, \quad \mathcal{L}^R = ||\ddot{M}_t^G - M^G||_1, \tag{7}$$

where M^G is the ground-truth monaural spectrogram for the target. Note that the predictions from f^B and f^M (i.e., $\tilde{B}_t^G$ and $\tilde{M}_t^G$ respectively) are step-wise predictions, unlike f^R that takes into consideration the history of monaural predictions in the episode to refine its estimate. Hence, we pretrain f^B and f^M using $\mathcal{L}^B$ and $\mathcal{L}^M$ and a dataset we collect from the training scenes. For each datapoint in this dataset, we place the agent and k audio sources randomly in the scene, then at the agent location we record the ground truth spectrograms (B^G, M^G) for a randomly sampled target type. We find this pretraining stage leads to higher performance and brings more stability to the predictions of the audio separator network compared to training those modules on-policy.

Once f^B and f^M are trained, we freeze their parameters and train f^R on-policy along with the audio-visual controller, since the sequence of actions taken by the agent impact the history of the monaural predictions observed by f^R.

Training the Active Audio-Visual Controller. The policy guides the agent to improve its audio separation quality by moving around. Towards this goal, we formulate a novel dense RL reward to train the quality policy $\pi^{\mathcal{Q}}$:

$$r_t = \begin{cases} r_t^s & 1 \leq t \leq \mathcal{T}-2 \\ -10 \times \mathcal{L}_{\mathcal{T}}^R + r_t^s & t = \mathcal{T}-1, \end{cases} \tag{8}$$

where $r_t^s = \mathcal{L}_t^R - \mathcal{L}_{t+1}^R$ is the step-wise reward that captures the improvement in separation quality of the monaural audio, and $r_{\mathcal{T}-1}$ is a one-time sparse reward at the end of the episode. While r_t^s encourages the agent to improve the separation quality at each step, the final reward $r_{\mathcal{T}-1}$ encourages the agent to take a trajectory that leads to an overall high-quality separation in the end. For the navigation policy $\pi^{\mathcal{N}}$, we adopt a typical navigation reward [54, 15, 16] and reward the agent with +1.0 for reducing the geodesic distance to the target source and an equivalent penalty for increasing it.

We train $\pi^{\mathcal{Q}}$ and $\pi^{\mathcal{N}}$ using Proximal Policy Optimization (PPO) [55] with trajectory rollouts of 20 steps. The PPO loss consists of a value network loss, policy network loss, and entropy loss to encourage exploration (see Supp.). Both policies have the same architecture but differ in their reward functions and distribution of their initial agent locations p_0.

Cyclic Training. We train the audio memory refiner f^R and the policies $\pi^{\mathcal{Q}}$ and $\pi^{\mathcal{N}}$ jointly. We adopt a cyclic training scheme, i.e., in each cycle, we alternate between training the audio memory refiner and the policy for $U = 6$ parameter updates. The cyclic training helps with stabilizing the RL training by ensuring partial stationarity of the rewards, particularly while training $\pi^{\mathcal{Q}}$, where the reward is a function of the quality of the target monaural predictions from f^R.

5. Experiments

Experimental Setup. For each episode, we place $k = 2$ (we also test with $k = 3$) audio sources randomly in the scene at least 8 m apart, and designate one as the target. The agent starts at the target audio location for the *near-target* task, and at a random location 4 to 12 m from other sources for the *far-target* task. The 12 m upper limit ensures that the agent can hear the target audio at the onset of its trajectory. We set the maximum episode length to $\mathcal{T} = 20$ and 100 steps for *near-target* and *far-target*, respectively. $\mathcal{T}^{\mathcal{N}} = 80$ is set using the validation split. We use all 47 Matterport3D scenes that are large enough to generate at least 500 distinct episodes given the setup above. We form train/val/test splits of 24/8/15 scenes and 112K/100/1K episodes. Because the test and train/val environments are disjoint, the agent is always tested in an unmapped space.

We use 12 types of sounds from three main groups: speech, music, and background sounds. For speech, we sample 10 distinct speakers from the VoxCeleb1 dataset [38] with different genders, accents, and languages. For music, we use a variety of instruments from the MUSIC dataset [77]. For background sounds, we sample non-speech and non-music sounds (e.g., clock-alarm, dog barking, washing machine) from ESC-50 [45]. The target in each episode can be one of $G \in \{Speakers, Music\}$, and the distractor(s) can be one of $D \in \{Speakers, Music, Background\}$ such that $G \neq D$ in the episode. Note that $Speakers$ denotes 10 separate speaker classes, resulting in a total of 11 target and 12 distractor classes. This enables us to evaluate a variety of audio separation scenarios: fine-grained separation (among the different speakers), coarse-grained separation (speech vs. music), and separation against background and ambi-

ent sounds commonly encountered in daily life. In total, we sample 23,677 1 sec audio clips of all types for use as monaural sounds. When testing on unheard sounds, we split the monaural sounds in the train:val:test ratio of 16:1:2. The longer audio clips used to produce the 1 sec unheard audio clips have no overlap among train, val, and test.

See Supp. for all other details like spectrograms, network architectures, training hyperparameters, and baseline details.

Baselines. Since no prior work addresses the proposed task, we design strong baselines representing policies from related tasks and passive/un-intelligent motion policies:

- **Stand In-Place:** audio-only baseline where the agent holds its starting pose for all steps, representing a default passive source separation method.
- **Rotate In-Place:** audio-only baseline where the agent stays at the starting location and keeps rotating in place, i.e., sampling acoustic cues from different orientations.
- **DoA:** Inspired by [40], this agent faces the audio direction of arrival (DoA), i.e., it directs its microphones at the target sound from one step away (only relevant for *near-target*).
- **Random:** an agent that randomly selects an action from the action space $\mathcal{A}$.
- **Proximity Prior:** an agent that selects random actions but stays within a radius of 2 m (selected via validation) of the target so it cannot wander far from locations that are likely better for separation. Note that this baseline assumes an oracle for distance to target, not given to our method.
- **Novelty [8]:** standard visual exploration agent trained to visit as many novel locations as possible within $\mathcal{T}$.
- **Audio-visual (AV) Navigator [15]:** state-of-the-art deep RL AudioGoal navigation agent [15] adapted for our task to additionally take the target audio category as input. Its audio input space exactly matches that of f^B and it is trained with the typical navigation reward [54, 15, 16].

For fair comparison, all baselines use our audio separator network f^A as the audio separation backbone, taking as input the audio/visual observations resulting from their chosen movements in the scene. Specifically, all agents share the same f^B and f^M, and only the audio memory refiner f^R is trained online with its respective policy. This means that any differences in performance are attributable to the quality of each method's action selection.

Evaluation. We evaluate the target monaural separation quality at the end of $\mathcal{T}$ steps, for 1000 test episodes with 3 random seeds. We use the ground-truth monaural phase [58] for all methods and the inverse short-time Fourier transform to reconstruct a time-discrete monaural waveform from $\ddot{M}^G$. We use standard metrics: **STFT** distance, a spectrogram-level measure of prediction error, and **SI-SDR** [52], a scale-invariant measure of distortion in the reconstructed signal.

Model	*Heard* SI-SDR ↑	*Heard* STFT ↓	*Unheard* SI-SDR ↑	*Unheard* STFT ↓
Stand In-Place	3.49	0.287	2.40	0.325
Rotate In-Place	3.45	0.285	2.50	0.321
DoA	3.63	0.280	2.59	0.316
Random	3.68	0.280	2.57	0.319
Proximity Prior	3.74	0.276	2.63	0.315
Novelty [8]	3.82	0.276	2.86	0.318
Move2Hear (Ours)	**4.31**	**0.260**	**3.20**	**0.298**

Table 1: Near-Target AAViSS[3].

Model	*Heard* SI-SDR ↑	*Heard* STFT ↓	*Unheard* SI-SDR ↑	*Unheard* STFT ↓
Stand In-Place	0.74	0.390	0.09	0.416
Rotate In-Place	1.01	0.382	0.26	0.412
Random	1.15	0.378	0.46	0.402
Novelty [8]	1.74	0.356	1.31	0.367
AV Navigator [15]	1.46	0.368	0.72	0.396
Move2Hear (Ours)	**3.50**	**0.291**	**2.33**	**0.333**

Table 2: Far-Target AAViSS.

5.1. Active Audio-Visual Source Separation

Near-Target. Table 1 reports the separation quality of all models on the *near-target* task.[3] Passive models that stay at the target (Stand, Rotate) do not perform as well as those that move (e.g., Proximity Prior). DoA fares better than the In-Place baselines as it gets to direct its microphones towards the target to sample a cleaner signal. Novelty outperforms the other baselines, showing the benefit of adding vision and sampling diverse acoustic cues. Our Move2Hear model outperforms all baselines by a statistically significant margin (according to the Kolmogorov–Smirnov test with $p \leq 0.05$). Move2Hear learns to take deliberate sequences of actions to improve the separation quality by reasoning about the 3D environment and inferred source locations.

Fig. 3a shows performance across each step in the episode. The non-stationary models make progress initially when sampling the cues close to the target, but then flatten quickly. In contrast, Move2Hear keeps improving with almost each action it takes, anticipating locations better for separation and learning behavior distinct from the other motion policies.

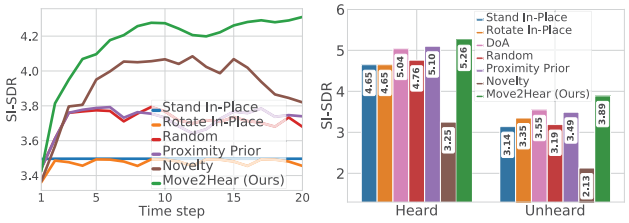
Far-Target. Table 2 shows the results on the *far-target* task. Here again we see a distinct advantage for models that move around. Interestingly, AV Navigator [15] performs worse than Novelty [8] even though it has been trained to navigate towards the target. This highlights the difficulty of audio goal navigation in the presence of distractor sounds and the need for high-quality separations for successful navigation. Our model outperforms the previous baselines by a significant margin ($p \leq 0.05$).

5.2. Model Analysis

[3] AV Navigator [15] is not applicable here; the agent begins at the target.

Model	*Heard* SI-SDR ↑	*Heard* STFT ↓	*Unheard* SI-SDR ↑	*Unheard* STFT ↓
Move2Hear	**3.50**	**0.291**	**2.33**	**0.333**
Move2Hear w/o f^R	2.64	0.320	1.57	0.361
Move2Hear w/o V_t	3.32	0.300	2.12	0.343
Move2Hear w/o $\pi^{\mathcal{N}}$	2.64	0.318	1.88	0.347
Move2Hear w/o $\pi^{\mathcal{Q}}$	3.08	0.304	1.99	0.343

Table 3: Ablation of our Move2Hear model on Far-Target AAViSS.



(a) Separation progress **(b)** Separation with 3 sources

Figure 3: (a) Separation quality as a function of time. (b) Final separation performance with 3 sources (i.e., 2 distractors).

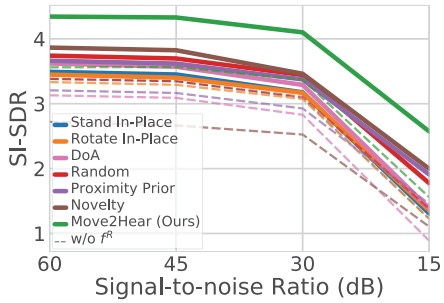


Figure 4: Models' robustness to various levels of noise in audio.

Ablations. In Table 3 we ablate the components of our model. We see that our acoustic memory refiner (f^R) plays an important role in the overall performance. f^R promotes stable, improved predictions by informing the policy of the separation quality changes. The vision component V_t is critical as well since V_t helps the agent to avoid obstacles, to reach the target, and to reason about the visible 3D scene.

Please see Supp. for **additional analysis** showing that 1) the composite policy for *far-target* is essential for best performance, 2) source types affect task difficulty, 3) Move2Hear retains its benefits even with a SOTA passive separation backbone, 4) Move2Hear's separation quality does not degrade with different intra-source distances, and 5) our model helps audio-visual navigation in the presence of distractor sounds.

Noisy Audio. We analyze our model's robustness to audio noise using standard noise models [64, 63] in the *heard* and *near-target* setting (Fig. 4). Our model's gains over the baselines persist even with increasing microphone noise levels. The plot also shows the positive impact of our memory refiner f^R (dashed lines); all models decline without it.[4]

[4]Note that evaluating noisy odometry and actuation is not supported by SoundSpaces since RIRs are available only on the discrete grid.

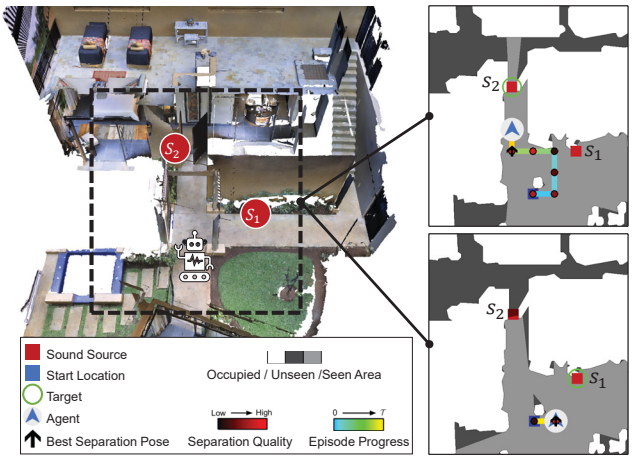


Figure 5: Example movements by our Move2Hear model. Our model takes advantage of the visible 3D structure to actively improve its separation quality of a target audio (see text for details).

Number of Audio Sources. Next we test how our model generalizes to more than one distractor sound. Fig. 3b shows the results for the *near-target* task using $k = 3$ audio sources per episode. Our model generalizes better than the rest of the baselines and maintains its advantage.

Qualitative Results. In Fig. 5, our Move2Hear agent is placed in a scene with two audio sources as possible targets. Our model exhibits an intriguing behavior that takes advantage of the visible 3D structures. When S_1 is the target, it takes the minimum steps to go around the column to put itself in the acoustic shadow of the column in relation to S_2, thus dampening its signal. However, when S_2 is the target, it decides to move into the corridor closer to S_2, putting the wall between itself and S_1.

Failure Cases. Common failure cases for *near-target* involve the agent having limited freedom of movement due to complex surrounding geometry and when any translational motion takes it towards the distractor(s) thus incurring a high loss in quality. For *far-target*, the agent sometimes lacks a direct path to the target due to cluttered surroundings.

6. Conclusion

We introduced the AAViSS task, where agents must move around using both sight and sound to best listen to a desired target object. Our Move2Hear model offers promising results, consistently outperforming alternative exploration/navigation motion policies from the literature, as well as strong baselines. In future work, we aim to extend our model to account for non-periodic sounds, e.g., with new forms of sequential memory, and to investigate sim2real transfer of the learned policies.

Acknowledgements: UT Austin is supported in part by DARPA L2M and the IFML NSF AI Institute. K.G. is paid as a Research Scientist by Facebook AI.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018. 1, 2

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. *arXiv preprint arXiv:1907.04975*, 2019. 2

[3] Xavier Alameda-Pineda and Radu Horaud. Vision-guided robot hearing. *The International Journal of Robotics Research*, 2015. 3

[4] Futoshi Asano, Masataka Goto, Katunobu Itou, and Hideki Asoh. Real-time sound source localization and separation system and its application to automatic speech recognition. In *Eurospeech*, 2001. 3

[5] Yutong Ban, Xiaofei Li, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Accounting for room acoustics in audio-visual multi-speaker tracking. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. 3

[6] Z. Barzelay and Y. Y. Schechner. Harmony in motion. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2

[7] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020. 2

[8] Marc G Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *arXiv preprint arXiv:1606.01868*, 2016. 7

[9] Gabriel Bustamante, Patrick Danes, Thomas Forgue, Ariel Podlubne, and Jérôme Manhès. An information based feedback control for audio-motor binaural localization. *Autonomous Robots*, 2018. 3

[10] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. MatterPort3D dataset license available at: http://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf. 2, 4

[11] Matthew Chang, Arjun Gupta, and Saurabh Gupta. Semantic visual navigation by watching youtube videos. In *Advances in Neural Information Processing Systems*, 2020. 2

[12] Devendra Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological slam for visual navigation. In *CVPR*, 2020. 2

[13] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *NeurIPS*, 2020. 2

[14] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *CVPR*, 2021. 1, 2, 3

[15] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. 1, 2, 3, 4, 6, 7

[16] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *International Conference on Learning Representations*, 2021. 1, 2, 3, 6, 7

[17] Soo-Whan Chung, Soyeon Choe, Joon Son Chung, and Hong-Goo Kang. Facefilter: Audio-visual speech separation using still images. *arXiv preprint arXiv:2005.07074*, 2020. 1, 2

[18] Antoine Deleforge and Radu Horaud. The Cocktail Party Robot: Sound Source Separation and Localisation with an Active Binaural Head. In *HRI 2012 - 7th ACM/IEEE International Conference on Human Robot Interaction*, pages 431–438, Boston, United States, Mar. 2012. ACM. 2

[19] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1830–1840, 2010. 2

[20] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 1, 2

[21] John W Fisher III, Trevor Darrell, William Freeman, and Paul Viola. Learning joint statistical models for audio-visual fusion and segregation. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 772–778. MIT Press, 2001. 2

[22] Jorge Fuentes-Pacheco, José Ruíz Ascencio, and Juan M. Rendón-Mancha. Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review*, 43:55–81, 2012. 2

[23] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. *arXiv preprint arXiv:1711.08789*, 2017. 1, 2

[24] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 2

[25] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020. 1, 2, 3

[26] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *ECCV*, 2020. 2, 3

[27] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 1, 2

[28] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888, 2019. 1, 2, 5

[29] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. *arXiv preprint arXiv:2101.03149*, 2021. 1, 2

[30] Rongzhi Gu, Lianwu Chen, Shi-Xiong Zhang, Jimeng Zheng, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu. Neural spatial filter: Target speaker speech separation assisted with directional information. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 4290–4294. ISCA, 2019. 2

[31] Rongzhi Gu and Yuexian Zou. Temporal-spatial neural filter: Direction informed end-to-end multi-channel target speech separation. *arXiv preprint arXiv:2001.00391*, 2020. 2

[32] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *CVPR*, 2017. 2, 4

[33] John R Hershey and Javier R Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *NeurIPS*, 2000. 2

[34] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Deep learning for monaural speech separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1562–1566, 2014. 2

[35] B. Li, K. Dinesh, Z. Duan, and G. Sharma. See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2906–2910, 2017. 2

[36] Dmytro Mishkin, Alexey Dosovitskiy, and Vladlen Koltun. Benchmarking classic and learned navigation in complex 3d environments. *arXiv preprint arXiv:1901.10915*, 2019. 2

[37] Arsalan Mousavian, Alexander Toshev, Marek Fišer, Jana Košecká, Ayzaan Wahid, and James Davidson. Visual representations for semantic target driven navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8846–8852. IEEE, 2019. 2

[38] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 6

[39] Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi G Okuno, and Hiroaki Kitano. Real-time speaker localization and speech separation by audio-visual integration. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, volume 1, pages 1043–1049. IEEE, 2002. 2

[40] Kazuhiro Nakadai, Tino Lourens, Hiroshi G Okuno, and Hiroaki Kitano. Active audition for humanoid. In *AAAI*, 2000. 3, 7

[41] Kazuhiro Nakadai and Keisuke Nakamura. Sound source localization and separation. *Wiley Encyclopedia of Electrical and Electronics Engineering*, 1999. 3

[42] Tsubasa Ochiai, Marc Delcroix, Yuma Koizumi, Hiroaki Ito, Keisuke Kinoshita, and Shoko Araki. Listen to what you want: Neural network-based universal sound selector. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 1441–1445. ISCA, 2020. 2

[43] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1, 2

[44] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Pérez, and G. Richard. Motion informed audio source separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6–10, 2017. 2

[45] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press. 6

[46] Jie Pu, Yannis Panagakis, Stavros Petridis, and Maja Pantic. Audio-visual object localization and separation using low-rank and sparsity. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2901–2905. IEEE, 2017. 2

[47] Senthil Purushwalkam, Sebastian Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. *arXiv preprint arXiv:2012.15470*, 2020. 2, 3

[48] Santhosh K. Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy Anticipation for Efficient Exploration and Navigation. In *European Conference on Computer Vision (ECCV)*, Aug. 2020. 4

[49] Omkar Ranadive, Grant Gasser, David Terpay, and P. Seetharaman. Otoworld: Towards learning to separate by learning to move. *ArXiv*, abs/2007.06123, 2020. 3

[50] Caleb Rascon and Ivan Meza. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 2017. 3

[51] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]). 4

[52] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey. Sdr – half-baked or well done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630, 2019. 7

[53] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation. *arXiv preprint arXiv:1803.00653*, 2018. 2, 4

[54] 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. 2, 4, 6, 7

[55] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 6

[56] F. Sedighin, M. Babaie-Zadeh, B. Rivet, and C. Jutten. Two multimodal approaches for single microphone source separation. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 110–114, 2016. 2

[57] P. Seetharaman, G. Wichern, S. Venkataramani, and J. L. Roux. Class-conditional embeddings for music source separation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 301–305, 2019. 2

[58] Andrew JR Simpson, Gerard Roma, and Mark D Plumbley. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *International Conference on Latent Variable Analysis and Signal Separation*,

pages 429–436. Springer, 2015. 7

[59] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In Mike E. Davies, Christopher J. James, Samer A. Abdallah, and Mark D. Plumbley, editors, *Independent Component Analysis and Signal Separation*, pages 414–421, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. 2

[60] Paris Smaragdis and Paris Smaragdis. Audio/visual independent components. In *in Proc. of International Symposium on Independant Component Analysis and Blind Source Separation*, 2003. 2

[61] Martin Spiertz and Volker Gnann. Source-filter based clustering for monaural blind source separation. In *in Proceedings of International Conference on Digital Audio Effects DAFx'09*, 2009. 2

[62] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2

[63] Ryu Takeda and Kazunori Komatani. Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2217–2221. IEEE, 2017. 8

[64] Ryu Takeda, Yoshiki Kudo, Kazuki Takashima, Yoshifumi Kitamura, and Kazunori Komatani. Unsupervised adaptation of neural networks for discriminative sound source localization with eliminative constraint. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3514–3518, 4 2018. 8

[65] Sebastian Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002. 2

[66] Efthymios Tzinis, Scott Wisdom, John R Hershey, Aren Jansen, and Daniel PW Ellis. Improving universal sound separation using sound classification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 96–100. IEEE, 2020. 2

[67] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020. 2

[68] Raquel Viciana-Abad, Rebeca Marfil, Jose Perez-Lorenzo, Juan Bandera, Adrian Romero-Garces, and Pedro Reche-Lopez. Audio-visual perception system for a humanoid robotic head. *Sensors*, 2014. 3

[69] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. On Audio, Speech and Lang. Processing*, 2007. 2

[70] Ron J Weiss, Michael I Mandel, and Daniel PW Ellis. Source separation based on binaural cues and source model constraints. In *Ninth Annual Conference of the International-Speech Communication Association, 2008*, 2009. 2

[71] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019. 2

[72] Yi Wu, Yuxin Wu, Aviv Tamar, Stuart Russell, Georgia Gkioxari, and Yuandong Tian. Bayesian relational memory for semantic visual navigation. In *ICCV*, 2019. 2

[73] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 2

[74] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 882–891, 2019. 2, 5

[75] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors, 2019. 2

[76] Xueliang Zhang and DeLiang Wang. Deep learning based binaural speech separation in reverberant environments. *IEEE/ACM transactions on audio, speech, and language processing*, 25(5):1075–1084, 2017. 2

[77] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 2, 5, 6

[78] 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. 2

[79] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *International Conference on Robotics and Automation (ICRA)*, pages 3357–3364, 2017. 2

[80] Özgür Yılmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. In *IEEE TRANSACTIONS ON SIGNAL PROCESSING (2002) SUBMITTED*, 2004. 2

[81] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký. Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):800–814, 2019. 2