Hearing to Hunt: Acoustic-Derived Observation Spaces for Combat Policy Learning

Jessenth Ebenezer Sankar[1]
js15339@nyu.edu

[1]**New York University**

Abstract

We investigate whether reinforcement learning agents can learn effective combat policies using spatial acoustic-derived observations, without visual input or privileged geometric state access. We introduce a controlled first-person combat environment in Unity and train agents using Proximal Policy Optimization (PPO) under two observation regimes: a privileged geometric baseline that computes spatial features directly from ground-truth game state as an upper-bound oracle, and a perceptual audio observer that derives equivalent features from Unity's spatial audio engine runtime. Through systematic feature ablation, we demonstrate that listener-relative azimuth is the dominant feature required for effective combat policy learning: its removal drives performance down to near-random levels regardless of reward shaping or training duration. Distance information proves to be largely redundant. Agents trained with the perceptual audio observer achieve performance comparable to the privileged baseline, with both conditions converging to mean episode rewards of 7–8, indicating consistent episode wins against the enemy in combat. These results suggest that engine-derived acoustic features provide sufficient perceptual grounding for combat policy learning, offering a practical alternative to visual or privileged-state observation spaces in game AI.

1 Introduction

Games have long served as the canonical testbed for reinforcement learning research, offering well-defined reward signals, controllable complexity, and reproducible evaluation conditions that the messy real world rarely provides. From early tabular methods on simple grid worlds (Sutton & Barto, 2018) to deep RL agents mastering Atari (Mnih et al., 2015) and defeating world champions at Go and StarCraft (Silver et al., 2016), games have driven the field forward precisely because they let us isolate what an agent can and cannot learn.

Yet, despite this tradition of careful experimental control, a persistent assumption has gone largely unexamined: that agents either receive rich visual input or, worse, direct access to privileged game state such as exact enemy coordinates, full map information, entity positions read straight from memory. The latter is a form of "wallhacking": the agent bypasses the perceptual constraints human players operate under, knowing where every entity is at all times without ever having to perceive them.

Human players, by contrast, rely heavily on non-visual cues. In competitive first-person games, experienced players track enemy positions through footstep sounds, gunshot direction, and reload audio. Spatial audio is not merely supplementary: it is a primary channel for situational awareness. This raises a natural question: *can a reinforcement learning agent learn effective combat behavior using an acoustic-derived observation channel in place of privileged state?*

We investigate this question by introducing a controlled, audio-only first-person combat environment in Unity. To systematically characterize the informational value of spatial audio, we define two observation regimes: a *privileged geometric baseline* using ground-truth game state as an upper-bound oracle, and a *perceptual audio observer* deriving features directly from the engine's simulated sound propagation. Our goal is not to demonstrate that acoustic-derived features contain more information than privileged geometry, but to investigate whether spatial acoustic-derived features alone constitute a sufficient perceptual interface for learning effective combat policies.

In this work, we: (1) introduce a novel combat environment constrained entirely to acoustic-derived directional and distance observations, augmented by a small set of non-acoustic proprioceptive and event flags shared identically across both regimes; (2) demonstrate through feature ablation that listener-relative azimuth is the critical feature for combat policy learning while distance is largely redundant; and (3) show that perceptual audio agents match baseline performance, offering a computationally viable alternative to privileged state access for audio-driven game AI.

2 Related Work

2.1 Audio in Reinforcement Learning

The use of audio as an observation modality in RL has received growing but limited attention. Hegde et al. (2021) extended VizDoom to support high-throughput audio-visual observation collection, demonstrating that combined audiovisual agents outperform vision-only agents in first-person duel scenarios. Chen et al. (2020) introduced SoundSpaces, an audio-visual navigation benchmark built on Habitat using geometrically simulated room impulse responses, showing that acoustic observations aid goal-directed navigation in realistic 3D environments. In both cases, audio serves as a complementary modality alongside vision. Giannakopoulos et al. (2021) trained a PPO agent on raw stereophonic audio for navigation and sound source localization in a Unity environment, demonstrating that audio-only policies are learnable for navigation tasks. Our work extends this line of research to adversarial combat, a substantially harder task requiring active target localization, approach, and precise timing under incoming fire.

2.2 Blind Agents in Games

Nguyen & Thawonmas (2022) introduced a blind deep RL agent for DareFightingICE, a fighting game platform with a sound-only observation mode, using PPO with audio encoders including 1D-CNN, FFT, and Mel-spectrogram applied to raw waveforms. While this establishes the feasibility of audio-only RL in games, the setting differs from ours in two key ways: first, fighting games provide structured, frame-synchronous sound events rather than continuous spatial audio requiring localization; second, their work does not ablate individual acoustic features to characterize which carry decision-relevant information. Our work directly addresses this gap by systematically ablating spatial acoustic features in a combat environment.

2.3 RL in First-Person Environments

VizDoom (Kempka et al., 2016) remains the standard platform for first-person RL research. While it supports stereo audio, its default audio subsystem is not optimized for faster-than-realtime training (Hegde et al., 2021). Our environment is purpose-built in Unity using ML-Agents (Juliani et al., 2018), with explicit control over spatial audio parameters, enemy behavior, and observation modalities, enabling clean ablation studies not easily achievable in existing platforms.

3 Background

3.1 Observation Spaces in Game AI

How an agent perceives its environment fundamentally determines what it can learn. In game AI research, observation spaces generally fall into three categories. *Raw pixel inputs* provide the agent with rendered frame buffers, typically processed through convolutional neural networks (Mnih et al., 2015). While expressive, this approach requires the agent to learn computer vision from scratch before any strategic reasoning can emerge, making it computationally expensive and sample-inefficient. *Raycasting* fires geometric rays outward from the agent, returning distances and object tags upon collision - a structured, lightweight approximation of vision that remains dependent on line-of-sight. *Privileged state access* bypasses perception entirely, providing exact entity coordinates and game state variables directly from the engine. This approach trains efficiently but eliminates partial observability, effectively giving the agent a wallhack.

Spatial acoustic observations occupy a distinct position in this landscape. Like raycasting, they are computationally lightweight and yield structured low-dimensional features. Unlike raycasting, they carry information about entities *beyond line-of-sight* - a property that raw pixels and raycasting both lack, at least in principle when occlusion, reverberation, or noise are modeled. Sound can propagate around obstacles and convey event information (footsteps, gunshots) that visual modalities may miss entirely. In the obstacle-free arena studied here, this potential is not exercised; we return to this point in Section 7.

3.2 Acoustic Observation Regimes

To systematically characterize the informational value of spatial audio for combat, we define two observation regimes. The *privileged geometric baseline* computes spatial features directly from ground-truth game state: azimuth via signed angle between the agent's forward vector and the vector to the enemy, and distance via Euclidean distance - representing an omniscient upper bound with no perceptual noise or attenuation. The *perceptual audio observer* derives equivalent features from Unity's spatial audio engine runtime: azimuth is computed in the agent's local coordinate frame using `InverseTransformDirection`, and distance is modulated by the engine's physical volume rolloff curve, introducing realistic attenuation without requiring raw waveform processing. Both observers share the same six-dimensional observation vector and identical network architecture, enabling direct comparison.

4 Environment

4.1 Arena

We implement a custom first-person combat environment in Unity 6 using the ML-Agents toolkit (Juliani et al., 2018). The arena consists of a flat enclosed rectangular floor (30×30 units) bounded by four walls, with no internal obstacles. Both the agent and enemy are represented as capsule-shaped rigid bodies. At the start of each episode, the agent spawns at a randomized position in the left half of the arena and the enemy in the right half, both with randomized initial orientations. This randomization prevents the agent from exploiting fixed spawn strategies and encourages generalization of acoustic policies across varied encounter geometries.

4.2 Enemy Behavior

The enemy follows a scripted policy: it selects a random movement direction every 2 seconds, moves at 3 units per second, and continuously rotates to face the agent. It fires a hitscan shot toward the agent every 4 seconds with a shot accuracy of 0.2 (an angular miss offset of up to $0.2 \times 45^\circ$ in both axes), dealing 25 damage per hit. This produces a challenging but beatable opponent that forces the agent to act under genuine combat pressure without requiring a learned enemy policy.

4.3 Agent Action Space

The agent operates with four discrete action branches: forward/backward movement (3 options), left/right strafing (3 options), left/right rotation (3 options), and shoot (2 options: hold fire or fire). Movement and strafing operate in world space at 5 units per second; rotation applies at 180° per second. Shooting triggers a hitscan raycast along the agent's forward vector with a cooldown of 0.5 seconds and a maximum range of 20 units, dealing 25 damage on hit. Both agent and enemy begin each episode with 100 HP.

4.4 Reward Structure

The reward function is designed to incentivize aggressive combat behavior:

- $+1.0$ per successful hit on the enemy
- $+5.0$ for killing the enemy (episode win)
- -0.5 for agent death (episode loss)
- -2.0 for episode timeout (30 seconds)
- -0.0005 per step to discourage passivity

Episodes terminate on kill, agent death, or timeout. The asymmetric kill/death reward ($+5$ vs -0.5) deliberately incentivizes aggression over survival, ensuring the agent does not learn passive avoidance strategies.

4.5 Observation Space

Both observation regimes produce a six-dimensional vector $\mathbf{o}_t \in \mathbb{R}^6$:

$$\mathbf{o}_t = [\alpha_t,\ d_t,\ f_t,\ g_t,\ s_t,\ h_t] \tag{1}$$

where $\alpha_t \in [-1, 1]$ is the normalized azimuth to the enemy, $d_t \in [0, 1]$ is the normalized distance, $f_t \in \{0, 1\}$ is a footstep flag, $g_t \in [0, 1]$ is a decaying gunshot flag, $s_t \in \{0, 1\}$ is the agent's own shot flag, and $h_t \in [0, 1]$ is the agent's normalized HP. The two regimes differ *only* in how α_t and d_t are computed; the remaining four components, including the proprioceptive HP signal h_t, are identical, non-acoustic, and shared across both conditions. We flag this explicitly: the comparison isolates the method of computing directional and distance information, not a strictly acoustic-versus-visual perceptual boundary.

4.6 Training Infrastructure

To accelerate training, 16 arena instances are parallelized within a single Unity scene for the privileged geometric baseline experiments. Due to Unity's constraint of a single `AudioListener` per scene, the perceptual audio observer is trained with a single environment instance. Both regimes are evaluated on total environment steps rather than wall-clock time to control for this asymmetry; we discuss its remaining limitations in Section 7.

5 Experiments

5.1 Training Setup

All agents are trained using Proximal Policy Optimization (PPO) (Schulman et al., 2017) via Unity ML-Agents (Juliani et al., 2018). The policy network consists of 3 fully connected hidden layers of 256 units each with observation normalization enabled. Training runs for up to 1×10^6 environment steps depending on the experimental condition, with a time scale of 20 (simulation runs at $20\times$ real

time). Key hyperparameters are listed in Table 1. All reported results are from single training runs per condition; we discuss the implications of this for statistical confidence in Section 7.

Table 1: PPO hyperparameters used across all experiments.

Hyperparameter	**Value**
Batch size	128
Buffer size	4096
Learning rate	3×10^{-4} (linear decay)
Entropy coefficient β	5×10^{-3}
Clip range ϵ	0.2
GAE λ	0.95
Epochs per update	3
Time horizon	128
Hidden units	256
Hidden layers	3
Observation normalization	Enabled

5.2 Experimental Conditions

We evaluate the following conditions:

Privileged Geometric Baseline. The agent observes ground-truth azimuth and distance computed directly from game state. Trained with 16 parallel arenas for up to 1 million steps.

Perceptual Audio Observer. The agent observes engine-derived azimuth and volume-attenuated distance from Unity's spatial audio runtime. Trained with a single arena for up to 1 million steps. Both regimes are compared on total environment steps to ensure a fair measure of sample efficiency.

Ablation: No Azimuth. The azimuth feature α_t is removed from the observation vector (5-dimensional input). Trained under both shaped and unshaped reward conditions to confirm findings are not reward artifacts.

Ablation: No Distance. The distance feature d_t is removed from the observation vector (5-dimensional input). Trained under shaped reward conditions.

Reward Shaping Analysis. To isolate the impact of reward design on policy emergence, we compare the event-driven reward structure defined in Section 4 against a densely shaped alternative. The dense condition includes supplementary rewards for proximity to the enemy and facing angle alignment. While these shaping signals are computed using ground-truth geometric state, they are applied strictly to the training signal and are never exposed to the agent's observation space $\mathbf{o}_t$, preserving the acoustic-derived constraint on the perceptual channel under test.

6 Results

6.1 Main Results

Figure 1 shows the mean episode reward over training steps for the privileged geometric baseline and the perceptual audio observer. Both conditions converge to mean episode rewards of 7–8, with the perceptual audio observer reaching stable performance by approximately 250k–300k steps, while the privileged geometric baseline requires approximately 450k–550k steps to achieve comparable stability, both indicating consistent episode wins.

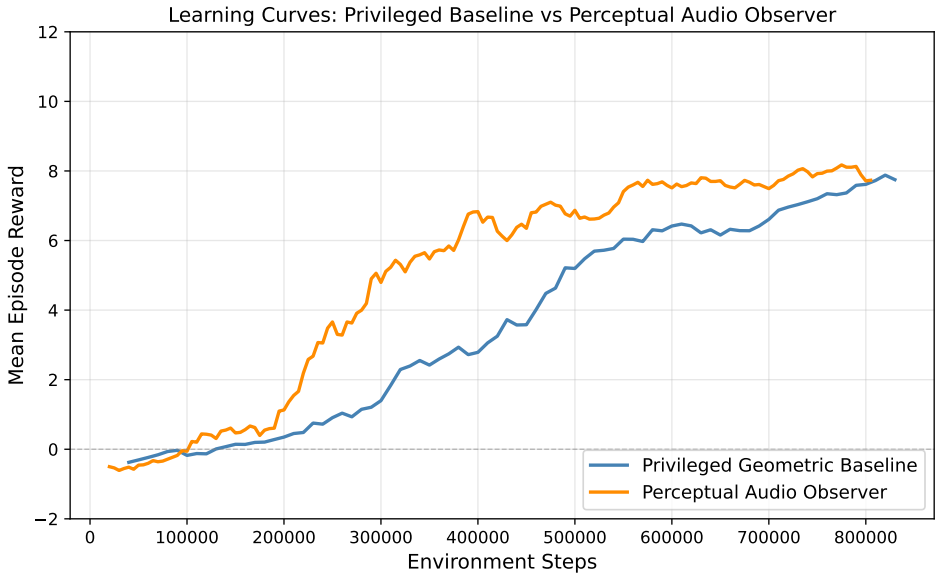

Figure 1: Mean episode reward over environment steps for the privileged geometric baseline and perceptual audio observer. Results from single training runs per condition. The perceptual audio observer converges earlier, reaching stable performance by approximately 250k–300k steps, while the privileged geometric baseline requires approximately 450k–550k steps.

6.2 Feature Ablation

Table 2 summarizes the performance of all experimental conditions at convergence. The no-azimuth ablation fails to converge under either reward condition: the shaped variant (-0.48) falls at or below the random-policy reference (-0.46), and the unshaped variant (0.81) remains far below the $\approx$ 7–8 range achieved by every azimuth-equipped condition, confirming that directional information is the dominant feature for combat policy learning. Removing distance information produces no degradation: the no-distance agent converges to a mean reward of 8.57, matching or exceeding both the privileged geometric baseline (7.66) and the perceptual audio observer (7.96).

Table 2: Mean episode reward at convergence for all experimental conditions. Results reported over the final 100k training steps.

Condition	**Obs. Dim.**	**Mean Reward**
Privileged geometric baseline	6	7.66
Perceptual audio observer	6	7.96
No azimuth (shaped)	5	-0.48
No azimuth (unshaped)	5	0.81
No distance (shaped)	5	8.57
Random policy (step 0)	6	≈ -0.46

7 Discussion

7.1 Azimuth as the Critical Feature

The most striking finding of this work is how far performance falls when azimuth is removed from the observation space. Under the shaped reward condition, the no-azimuth agent converges to -0.48,

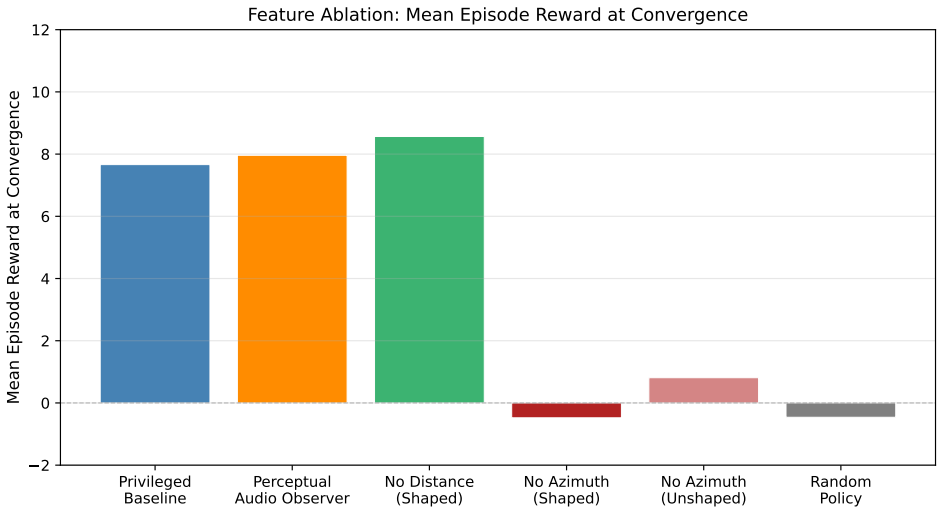

Figure 2: Mean episode reward at convergence across all experimental conditions, computed over the final 100k training steps of a single run per condition (n=1; bars shown without error bars, as no across-run variance is estimated). The privileged geometric baseline (7.66) and perceptual audio observer (7.96) converge to comparable performance, while removing distance produces no degradation (8.57). Removing azimuth drives performance down to near-random levels under both shaped (-0.48, at the random-policy reference of -0.46) and unshaped (0.81, still far below the 7–8 range of every azimuth-equipped condition) reward conditions, confirming azimuth as the dominant feature independent of reward design. The random-policy reference (-0.46) is the reward at the earliest logged step.

at or below the random-policy reference of -0.46 - indistinguishable from an agent that has learned nothing. Under the unshaped condition it reaches 0.81, still far below the $\approx$ 7–8 range achieved by every azimuth-equipped condition. Neither variant learns effective combat behavior. This suggests that directional information is not merely useful but necessary for combat policy learning - without knowing which direction to face, the agent cannot align its forward vector to shoot, and the task becomes effectively unsolvable.

By contrast, removing distance produces no measurable degradation at all: the no-distance agent converges to a mean reward of 8.57, exceeding both the privileged baseline (7.66) and the perceptual audio observer (7.96). We hypothesize that distance information is implicitly recoverable from the reward signal: an agent that faces the enemy and shoots will receive hit rewards only when close enough, providing an indirect distance signal through the reward structure itself. This redundancy is also plausibly specific to our task: with a hitscan weapon, a small arena, and reward feedback tied directly to successful hits, explicit range estimation may carry less value than it would in environments with projectile travel time, ranged-versus-melee tradeoffs, or larger traversal costs. Azimuth, however, has no such implicit substitute, as the agent must receive directional information explicitly to orient itself. We note the no-distance condition exceeding the six-feature baselines is itself somewhat counterintuitive, and may indicate that the removed distance dimension occasionally acted as a mildly distracting input for the policy network in this particular task, rather than a purely inert one.

7.2 Perceptual Audio vs. Privileged Baseline

The perceptual audio observer achieves mean episode rewards of 7–8, matching the asymptotic performance of the privileged baseline. Notably, the acoustic agent converges approximately 150k–200k steps earlier than the geometric baseline. We hypothesize that this accelerated sample efficiency stems from the structure of the engine-derived features: Unity's logarithmic volume rolloff

provides a natural, non-linear proximity gradient, while local-frame acoustic azimuth directly maps to the agent's egocentric rotational action space, bypassing the need to learn coordinate transformations from global geometric state.

This result should be read carefully. In our obstacle-free arena, Unity's spatial audio engine computes azimuth and distance from the same underlying transform data available to the privileged baseline; there is no occlusion, reverberation, or environmental noise to make the acoustic channel informationally distinct from geometry. What we demonstrate, precisely, is that routing the identical underlying information through the engine's audio subsystem rather than reading it directly from game state does not degrade policy learning, and if anything provides a more learnable feature encoding. We do not claim that the acoustic channel here contains information unavailable to the geometric baseline, only that the two are architecturally and empirically equivalent under this environment's conditions. Whether this equivalence holds once occlusion or ambient noise decouple acoustic signal from ground-truth geometry is an open question we do not answer here.

7.3 Reward Shaping

Dense reward shaping dramatically accelerates policy emergence. The shaped agent discovers effective combat behavior by 10k steps, compared to 230k steps for the unshaped agent. However, asymptotic performance is comparable in both conditions, confirming that the learned behavior is not an artifact of the shaping signal. This finding has practical implications: reward shaping using privileged geometric state can safely bootstrap acoustic agents without compromising the validity of the learned policy, provided the shaping signal is excluded from the observation space.

7.4 Limitations and Future Work

Several limitations of this work bound its conclusions and suggest directions for future research.

First, all reported results are drawn from single training runs per condition, without seed-averaged means, confidence intervals, or the statistical treatment recommended by Agarwal et al. (2021) for deep RL evaluation. The reported trends should be read as suggestive rather than statistically established, and replication across seeds is the most direct next step for strengthening these claims.

Second, the arena contains no obstacles, meaning sound propagation is unobstructed and occlusion, reverberation, and ambient noise are entirely absent. As a direct consequence, the acoustic-derived azimuth and distance features used here are recoverable from the same underlying geometric state as the privileged baseline; nothing about our arena forces a genuine perceptual distinction between hearing and knowing. Introducing walls, cover, and noise would create a setting where acoustic and geometric information can diverge, which is the setting in which claims of a uniquely acoustic advantage could be meaningfully tested.

Third, the training infrastructure differs between conditions: the privileged baseline uses 16 parallelized arenas while the audio observer, constrained by Unity's single-`AudioListener`-per-scene limitation, uses one. We control for this by comparing on total environment steps rather than wall-clock time, but this does not fully equalize the two setups, and residual differences in gradient batching or update variance between the two training regimes cannot be ruled out.

Fourth, the reward structure - particularly the reliance on a hitscan weapon and small arena - may make distance information less critical than it would be in settings with ranged combat, projectile travel time, or larger traversal distances; the finding that distance is redundant should be understood as specific to this task rather than a general property of combat policy learning.

Fifth, the scripted enemy follows a fixed behavioral policy; replacing it with a learned adversary would introduce non-stationarity and require more robust tracking of directional information over time.

Finally, while our perceptual audio observer avoids raw waveform processing, future work could investigate whether end-to-end training from raw stereo audio, using learned audio encoders such

as 1D-CNNs or spectrograms as in Nguyen & Thawonmas (2022), yields comparable or superior performance to our feature-engineered approach, and whether it forces the same or different conclusions about which acoustic cues matter.

8 Conclusion

We investigated whether reinforcement learning agents can learn effective combat policies from acoustic-derived spatial observations, without visual input or privileged geometric state access. Our results demonstrate that this is possible: engine-derived acoustic features match the performance of an omniscient privileged baseline, with both conditions converging to mean episode rewards of 7–8 under identical training conditions.

The central finding of this work is that azimuth, the directional angle to the enemy, is the dominant feature for combat policy learning regardless of whether it is computed from raw game state or routed through Unity's spatial audio engine. Its removal drives performance down to near-random levels regardless of reward design or training duration, while distance information proves largely redundant in this task. This asymmetry suggests that directional perception, not range estimation, is the fundamental requirement for combat policy learning in our setting.

We further demonstrate that reward shaping accelerates convergence dramatically without altering asymptotic policy quality, establishing that the learned behavior is not an artifact of reward design. Together, these findings characterize the minimal perceptual requirements for combat policy learning in an obstacle-free environment, and establish a testbed and methodology for future work that introduces the occlusion, noise, and multi-seed statistical rigor needed to test whether acoustic and geometric information genuinely diverge.

Broader Impact Statement

This work studies combat policy learning in a synthetic game environment and does not involve human subjects, personal data, or deployed systems. We do not anticipate direct negative societal impact from this research; potential downstream considerations are the same as those broadly applicable to game AI and simulated combat agents generally.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34:29304–29320, 2021.

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 *European Conference on Computer Vision*, pp. 17–36. Springer, 2020.

Petros Giannakopoulos, Aggelos Pikrakis, and Yiannis Cotronis. A deep reinforcement learning approach for audio-based navigation and audio source localization in multi-speaker environments. *arXiv preprint arXiv:2110.12778*, 2021.

Shashank Hegde, Anssi Kanervisto, and Aleksei Petrenko. Agents that listen: High-throughput reinforcement learning with multiple sensory systems. In *IEEE Conference on Games*, pp. 1–5. IEEE, 2021.

Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, et al. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*, 2018.

Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. ViZDoom: A Doom-based AI research platform for visual reinforcement learning. In *IEEE Conference on Computational Intelligence and Games*, pp. 341–348. IEEE, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Van Thai Nguyen and Ruck Thawonmas. A deep reinforcement learning blind AI in DareFightingICE. In *IEEE Conference on Games*, pp. 1–4. IEEE, 2022.

John Schulman, Filip Wolski, Prafull Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.