WavefrontNet: Sound-Based Navigation Agent

Danyil Butkovskyi[1]   Daniel Pitzele[1]

1. Introduction

Robust navigation in unstructured environments remains a fundamental challenge in robotics. While visual, LiDAR, and depth-based approaches have achieved significant success, they rely heavily on line-of-sight visibility and stable lighting conditions. In scenarios where these modalities are compromised—such as smoke-filled rooms, dark subterranean caves, or dense fog—traditional navigation agents become effectively blind.

In contrast, acoustic signals possess unique propagation properties that allow them to ”see” around corners and through visual occlusions. The way sound reverberates through a scene encodes rich information about the surrounding geometry, material composition, and free space. However, the use of passive audio for geometric inference and path planning remains significantly underexplored compared to visual modalities.

In this work, we propose that the spatial impulse response of a scene can be treated as a navigable feature map. We introduce **WavefrontNet**, a learning-based agent that navigates unmapped, complex 2D environments using only passive acoustic signals. Unlike active sonar, which requires emitting detectable pulses, our approach leverages ambient or goal-oriented sound sources to infer valid trajectories. By treating raw multi-channel audio as a spatio-temporal image, our model learns to interpret wavefront arrival patterns to effectively navigate toward a target, even in the absence of all visual input.

2. Background and Related Works

2.1. Navigation/Mapping

Prior to modern robotics, mapping subterranean environments was primarily driven by geological and archaeological interests [7, 11]. These early approaches relied on manual sensing to reconstruct natural phenomena. However, they produced static, explicit map representations that are often computationally expensive to maintain and update in real-time

Our work is different from previous mapping projects. Firstly, many of these mapping for navigation projects were conducted before the advent of machine learning. Therefore, they have little choice but to use an explicit representation of their constructed map as their output. Our project does not at all create an explicit representation of the environment. Instead, we allow the computational model to learn whatever patterns in the data it sees fit to improve navigation abilities, a unique approach from classical mapping works.

2.2. Robot Navigation

Recent advances in robot learning have allowed for robots to excel at tasks they never have previously. For example, Physical Intelligence is creating a generalist robot policy [2, 5] for household tasks such as folding laundry and washing dishes. Navigation is a crucial subtask in these daily, useful tasks since nearly all our daily tasks require movement around a scene.

Robotic navigation has also been explored in a similar context. [3, 4] show a closely related project that simulates autonomous robot navigation in pre-rendered 3D environments. They utilize a combination of visual, audio, and GPS signals in order to improve the robot’s ability to successfully map and navigate the scene. Although similar in kind to our project, there are notable differences between the two. Their project focused mainly on the end result of improved navigation abilities by a combination of many different modalities. On the other hand, we aim to showcase the unique ability of audio signals to encode certain aspects of a physical scene.

3. System Design

3.1. Environment/Dataset

We chose to study the case of an autonomous agent attempting to navigate an unmapped, cave-like 2D structure. The agent would be navigating to a set goal point that it was not aware of, but the goal point was constantly emitting a sound signal that was audible from the entire cave (at varying magnitudes). The agent had five possible moves: it could go left, right, up, or down on the 2D grid, or it could output ”stop” when it thought it was at the goal point. In order to solve this problem using machine learning techniques, we first needed to construct a dataset to train our model.

The first step to gathering data was cave generation. We

[1] Equal contribution, alphabetical order

needed to reliably, yet randomly, generate cave-like two-dimensional structures that had a few unique properties. Firstly, every position needed to be reachable from every other position to prevent the agent from getting stuck somewhere. Secondly, each pathway in the cave had to be sufficiently wide for the agent and its sensors to pass through. Finally, the cave should have a complex enough shape to challenge the navigation agent, but simple enough to be solvable with our relatively small training dataset. In order to generate this maze, we used a combination of graph-based algorithms and post-processing to ensure the caves had the desired properties.

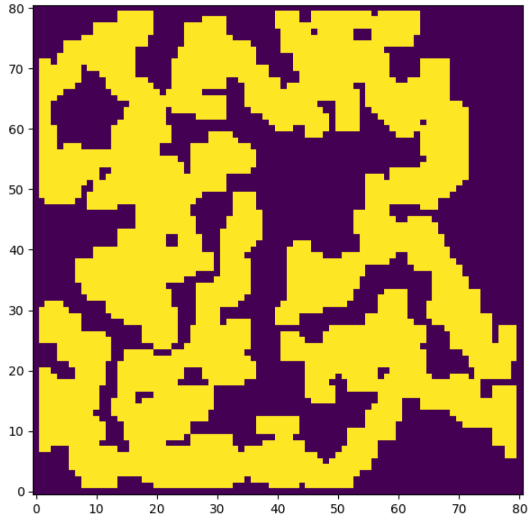


Figure 1. A randomly generated cave

Once our caves were generated, we faced the challenge of accurately simulating audio signals. We did this using a software called k-Wave [12], a wave signal simulation library in Python. k-Wave allowed us to simulate the propagation of the audio waves from the goal point in our arbitrary scenes, in this case the caves we had generated. Using this library, we obtained a fine-grained map (down to the centimeter) of the raw audio waveforms in the cave. We precomputed the audio responses for our caves in order to speed up online model training.

We sampled the audio waveforms on a grid in order to have precomputed values for any position that the agent could hold. As input, the agent would receive a 3x3 grid of raw audio waveforms around its position. We chose to use a grid so the agent could learn the correspondences between spatial variation in audio signals and the geometry of its scene. We chose to use raw waveforms as opposed to a post-processed version of the data in order to allow the model to learn whatever patterns it saw fit for the task.

In order to conduct supervised training, we needed to determine the ”correct” move for the agent at any given position. This was done using an A*-based pathfinding algorithm that starts at the goal point and creates a map of optimal directions. A known limitation of using naive A* in grid environments is that the algorithm often arbitrarily favors specific directions (e.g., zigzagging horizontally before vertically) when multiple paths are equally optimal. Ideally, this would be mitigated by calculating a **set** of valid moves for every position, treating all optimal directions as equally correct.

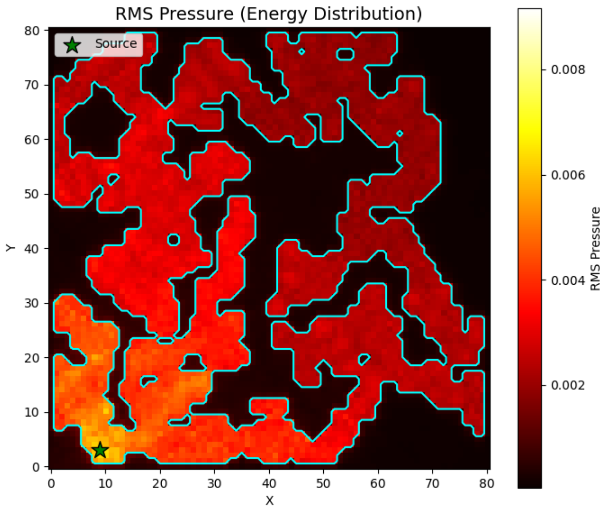


Figure 2. The pre-computed audio responses at each position in the cave

However, for the experiments presented in this work, we utilized a dataset where each position was assigned a single deterministic ground-truth label derived from the primary A* path. Although we acknowledge that this introduces some directional bias, requiring us to balance the training data by downsampling dominant classes, it provided a sufficient baseline to evaluate the core feasibility of acoustic navigation. Future iterations of the dataset generation will implement the multi-label approach to fully resolve these ambiguity artifacts.

Using these precomputed caves, audio waveforms, and deterministic path labels, we created our final training set. We conducted supervised training by sampling random positions in the caves, feeding the model audio waveforms from those positions, and backpropagating using the generated ground truth. The pre-computed nature of our dataset allowed us to reuse each maze structure efficiently across many training iterations.

3.2. Model Architecture

Traditional approaches to acoustic direction finding often rely on converting raw audio into time-frequency representations, such as Short-Time Fourier Transform (STFT)

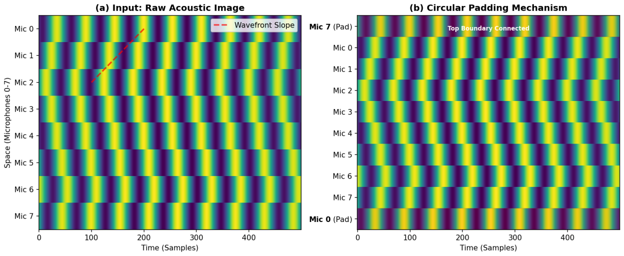


Figure 3. **The Spatio-Temporal Acoustic Image.** (a) Visualization of the raw 8-channel input. A sound source creates a visible diagonal wavefront (highlighted by the red dashed line); the slope of this line indicates the direction of arrival. (b) The Circular Padding mechanism. To respect the cylindrical topology of the physical array, the top row (Mic 7) is padded below Mic 0, and the bottom row (Mic 0) is padded above Mic 7.

spectrograms [1]. However, these magnitude-based transformations often discard or obscure crucial inter-channel phase information required for fine-grained localization [9].

To address this limitation, we propose **WavefrontNet**, an end-to-end 2D Convolutional Neural Network (CNN) that operates directly on raw multichannel waveforms. Our approach is inspired by recent successes in processing raw audio directly for speaker recognition and localization tasks [6, 8].

3.2.1. The Acoustic Image Representation

Instead of treating the input as eight independent 1D time-series, we stack the synchronized raw audio data into a unified 2D tensor, creating a single-channel "acoustic image" with dimensions $1 \times 8 \times 11,434$ (Channels × Space × Time).

As visualized in Figure 3(a), the physics of sound propagation creates a distinct visual pattern in this representation. A sound wave arriving from a specific direction does not manifest as a simple delay, but rather as a continuous **diagonal wavefront** across the image. The slope of this wavefront is directly proportional to the inter-channel time difference (ITD), encoding the angle of arrival. By framing the problem this way, acoustic localization is re-tasked as a visual edge-detection problem.

3.2.2. Geometric Adaptation via Circular Padding

A fundamental challenge in applying standard 2D CNNs to this data is the geometric mismatch. Standard convolutions assume planar data with distinct boundaries. However, our agent utilizes a Uniform Circular Array (UCA), where Microphone 0 is physically adjacent to Microphone 7 [10]. Standard zero-padding would introduce artificial edge artifacts, breaking the continuity of wavefronts arriving from directions between Mic 7 and Mic 0.

To resolve this topology mismatch, we implement a **Circular Padding** mechanism on the spatial dimension prior to convolution operations. As illustrated in Figure 3(b), this process logically connects the top and bottom boundaries of the input tensor. This topological transformation effectively turns the flat acoustic image into a cylinder, allowing convolutional kernels to track phase shifts continuously as they wrap around the physical array.

3.2.3. Feature Extraction Backbone

The WavefrontNet backbone consists of four convolutional blocks designed to detect directional slopes.

The first layer utilizes anisotropic kernels of size (3×64) with a stride of $(1, 4)$. The spatial kernel size of 3 allows the network to compare a central microphone with its immediate left and right neighbors simultaneously, creating a learnable local phase detector. The large temporal kernel size integrates information over ≈3ms, providing robustness against high-frequency noise.

To handle the high sampling rate (22.05 kHz) while preserving spatial fidelity, subsequent layers utilize **asymmetric Max Pooling**. We employ a pooling kernel size of (2×4) across the spatial and temporal dimensions, respectively. This strategy compresses the temporal dimension aggressively (reducing 11,434 samples to a scalar) while preserving spatial resolution deep into the network, ensuring that fine-grained directional cues are maintained until the final classification head.

4. Evaluation

4.1. Quantitative Results

We evaluated WavefrontNet on a held-out validation set comprising 19,572 samples from unseen cave environments. The model achieved a top-1 accuracy of **61.6%**. Given the four-way classification task (random baseline = 25%), this performance confirms that the model extracts robust directional cues from raw reverberant waveforms.

Table 1 details the per-class performance metrics. A striking asymmetry is observed between axes: the model performs significantly better on the horizontal axis (F1-Scores ≈ 0.70) than the vertical axis (F1-Scores ≈ 0.54). This suggests the acoustic features encoding "Left/Right" are more distinct than those encoding "Up/Down" in our current representation.

4.2. Error Analysis

The confusion matrix (Figure 4) reveals a crucial geometric insight.

For the horizontal axis, the model is highly robust: "reversal errors" are rare, with only 5.3% of *Left* samples misclassified as *Right*. However, on the vertical axis, confusion between opposites is the primary failure mode. For example, 21.7% of *Down* samples are incorrectly predicted as *Up*.

Table 1. Per-class performance metrics for WavefrontNet on the validation set.

Class	Precision	Recall	F1-Score	Support
UP	0.53	0.56	0.55	4893
DOWN	0.56	0.50	0.53	4893
LEFT	0.67	0.71	0.69	4893
RIGHT	0.71	0.69	0.70	4893
Overall / Avg	**0.62**	**0.62**	**0.61**	**19572**

This vertical ambiguity implies that the ”diagonal wavefront” signature for Up and Down trajectories may look nearly identical in the current acoustic image representation, likely due to symmetries in the 8-microphone circular array that the current resolution cannot fully resolve. Future work will investigate non-uniform array spacing to break this symmetry.

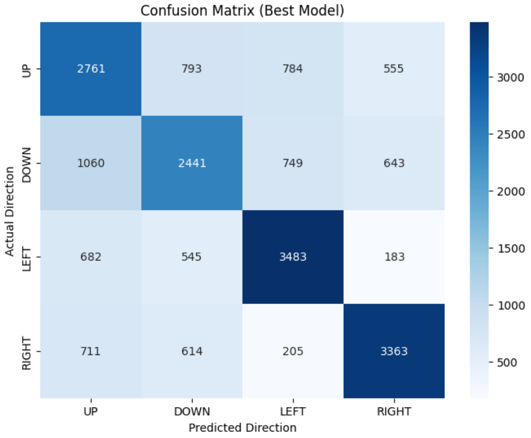
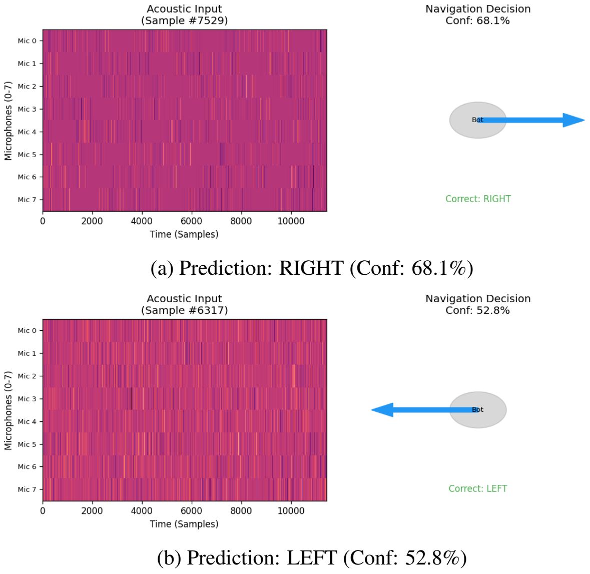


Figure 4. Confusion Matrix for WavefrontNet. While Left/Right separation is distinct, there is significant confusion between Up and Down, indicating a vertical symmetry ambiguity in the acoustic features.

4.3. Qualitative Analysis

To validate the model’s decision-making in successful cases, we visualize the model’s prediction alongside the raw input in Figure 5.

Figure 5(a) shows a sample where the target is **RIGHT**. The acoustic image displays a clear downward-slanting texture, which the model correctly identifies with 68.1% confidence. Figure 5(b) shows a **LEFT** target. Despite reverberation noise creating vertical smearing, the model detects the upward-slanting leading edge and correctly predicts the direction.



(a) Prediction: RIGHT (Conf: 68.1%)

(b) Prediction: LEFT (Conf: 52.8%)

Figure 5. **Live Navigation Decisions.** The left panel of each sample shows the raw ”Acoustic Image” input. The right panel shows the agent’s predicted vector (Blue arrow) matching the ground truth (Green text).

4.4. Training Setup

We trained the model for 100 epochs using the AdamW optimizer with an initial learning rate of 10^{-3} and a weight decay of 0.1. A “Reduce On Plateau” scheduler was employed to anneal the learning rate when validation performance plateaued.

To prevent overfitting on the synthetic dataset, we applied a rigorous data augmentation pipeline during training:

1. **Stochastic Time Shifting:** Randomly rolling the audio tensor along the temporal axis (± 1500 samples) to force the model to learn relative phase shifts rather than absolute positions.
2. **Noise Injection:** Injecting Gaussian noise scaled to 5% of the signal’s standard deviation.
3. **Label Smoothing:** Using a smoothing factor of 0.1 to prevent over-confidence in the classification head.

5. Conclusion and Future Work

In this work, we presented **WavefrontNet**, a novel end-to-end architecture for blind acoustic navigation. By reinterpreting multichannel audio as a spatio-temporal image and introducing a circular padding mechanism, we demonstrated that a robot can successfully navigate unmapped environments using only sound. Our model achieved 61.6% accuracy on a challenging four-way classification task, learning to identify directional wavefronts directly from raw audio without explicit feature engineering.

While these results establish the feasibility of the approach, several avenues remain for future exploration:

- **Continuous Angle of Arrival (AoA):** Currently, our agent is limited to four discrete moves (Up, Down, Left, Right). Future iterations will regress a continuous angle of arrival ($\theta \in [0, 2\pi]$) directly from the wavefront slope, enabling smoother, more natural control.
- **3D Navigation:** Our current simulation is restricted to 2D planar caves. Extending the acoustic image representation to handle elevation (using a spherical or cylindrical microphone array) would allow for fully 3D navigation in drones or underwater vehicles.
- **High-Fidelity Simulation:** While k-Wave provides accurate wave physics, our current environment assumes static, simplified wall impedance. Incorporating frequency-dependent absorption materials and dynamic sound sources would significantly reduce the ”sim-to-real” gap for physical deployment.
- **Ambiguity Resolution:** The error analysis revealed significant confusion between opposite directions on the vertical axis. Investigating non-uniform array geometries or adding recurrent memory units (LSTMs) could help the agent resolve these symmetries by integrating temporal context over multiple steps.

We believe that acoustic sensing is an underutilized modality in modern robotics. WavefrontNet represents a step toward agents that can ”see” with sound, opening new possibilities for operation in dark, foggy, or visually occluded environments.

References

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2019. 3

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. 1

[3] Changan Chen, Unnat Jain, Carl Schissler, Sebastià Vicenc Amengual Garí, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip W. Robinson, and Kristen Grauman. Audio-visual embodied navigation. *CoRR*, abs/1912.11474, 2019. 1

[4] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning, 2023. 1

[5] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. 1

[6] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028, 2018. 3

[7] W.I. Sellers and A.T. Chamberlain. Ultrasonic cave mapping. *Journal of Archaeological Science*, 25(9):867–873, 1998. 1

[8] Harshavardhan Sundar, Hualin Wang, Meng Wang, and Kun Wang. Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4642–4646, 2020. 3

[9] Paolo Vecchiotti, Ning Ma, Stefano Squartini, and Aurelio Uncini. End-to-end binaural sound localisation from the raw waveform. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 451–455, 2019. 3

[10] Mati Wax and Jacob Sheinvald. Direction finding of coherent signals via spatial smoothing for uniform circular arrays. *IEEE Transactions on Antennas and Propagation*, 42 (5):613–620, 1994. 3

[11] Nick Weidner, Sharmin Rahman, Alberto Quattrini Li, and Ioannis Rekleitis. Underwater cave mapping using stereo vision. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5709–5715, 2017. 1

[12] Farid Yagubbbayli, David Sinden, and Walter Simson. k-Wave-Python. 2