2022 IEEE International Conference on Robotics and Automation (ICRA)
May 23-27, 2022. Philadelphia, PA, USA

Learning to Listen and Move: An Implementation of Audio-Aware Mobile Robot Navigation in Complex Indoor Environment

Pratyaksh P. Rao[1], *IEEE Student Member* and Abhra Roy Chowdhury[2], *IEEE Senior Member*

***Abstract*— Sound is an essential navigation cue that intelligent robots can leverage for localizing sound-emitting targets. This work introduces a framework for the audio-aware navigation task of mobile robots equipped with a microphone array in a complex indoor environment. The robot initialized at a random starting position has to accurately localize a distant sound source and plan an optimal path towards the sound-emitting target. Auto-encoders are used to extract implicit acoustic features that are robust against environmental noise and reverberation. The proposed framework is based on two key ideas - a sound inference module (SIM) that maps the perceived acoustic information to a given geometric map of the physical space, and a path planner that generates a path from the robot's current position to the estimated position of the sound source. Experimental results show that the SIM achieved a minimum and maximum localization error of 0.31 m and 0.70 m at a robot-source distance of 1 m and 6 m, respectively at different environmental configurations. Additionally, the proposed framework achieved a minimum and maximum reliability of $4.38\ m^{-1}$ and $2.31\ m^{-1}$ at a robot-source distance of 1 m and 6 m, respectively under the influence of background noise.**

***Index Terms*— Audio-Aware Navigation, Sound Source Localization (SSL), Auto-Encoders, Convolutional Neural Networks (CNNs)**

I. INTRODUCTION

Sound is of paramount importance in understanding the world around us. It is an essential navigational cue utilized by many organisms. For instance, the larvae of reel fishes use acoustic signals to locate suitable settlement habitats [1]. Moreover, many animals leverage the sounds of other individuals to localize distant predators and prey. Furthermore, cognitive science confirms that audio signals provide vital information for navigation [2]. The navigation task of mobile robots has been studied extensively in the past. Traditional methods involve building a map of the environment and planning an optimal path within the same [3]. More recently, deep learning-based models are used to learn navigation policies and spatial representations from raw visual data to generalize in unseen environments [4], [5], [6], [7], [8]. However, these approaches focus mainly on visual perception and are deaf to the world around them. Therefore, this gives rise to a sensory privation, and such vision-dependent agents cannot detect sound-emitting targets. Recent studies on the embodied audio-visual navigation task [9], [10], [25] focus on integrating visual and auditory information for navigating through an unseen environment to locate a sound source. Methods in [26] explore how audio-visual cues can aid in floorplan reconstruction from limited viewpoints. The essence of this work is an open question: Can robots rely on audio cues for navigation in a real-world environment?

Fig. 1: An illustration of the experimental testbed.

Existing studies on audio-based navigation have two key limitations. Firstly, the audio signal recorded by the robot is a superposition of many sound sources present in the physical space. Therefore, the robot must retrieve the sound of interest and discard the rest from this mixture. Secondly, it is very challenging to conduct thorough and controlled experiments with physical robots. As a result, there is a gap between simulation and real-world implementation. In light of these limitations, this study investigates the possibility of adopting Deep Neural Networks (DNNs) that are capable of modeling complex associations between acoustic cues from the environment and geometric information from the constructed map, for the audio-aware navigation task. The task involves two crucial steps - 1) Sound Source Localization (SSL) and 2) Path planning. SSL is challenging due to the complexity and variability of the environment. Traditional techniques involve extracting explicit features from the audio signal such as time-difference-of-arrival (TDOA) or inter-microphone intensity difference (IID). These features do not provide adequate information about distant audio signals present in a noisy environment. In contrast, modern approaches that utilize DNNs are robust against background noise and environmental reverberation.

Fig. 1 illustrates the experimental testbed, where a mobile robot, equipped with a microphone array and depth camera,

[1]Pratyaksh P. Rao is with the Department of Electrical and Computer Engineering, New York University, New York, USA
[2]Prof. Abhra Roy Chowdhury is with the Centre for Product Design and Manufacturing (CPDM), Indian Institute of Science (IISc), Bangalore, India

2022 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9681-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRA46639.2022.9812193

must autonomously navigate to a distant sound source. It is assumed that a map of the environment is known to the robot, and it must utilize this map for estimating its pose in the environment. The robot is initialized at a random starting location and records the audio signal. The framework processes the acoustic information and localizes the sound source. Finally, a path planner generates a trajectory from its current location to the sound-emitting target. Advancements in this task will facilitate many applications in the industry.

This paper makes the following contributions:

- A framework is presented for the **Audio-Aware Navigation** of mobile robot. The proposed methodology is implemented on a physical system, validated in a complex manufacturing environment, and contrasted with other baselines.
- A **Sound Inference Module (SIM)** is introduced for high resolution **Sound Source Localization (SSL)**. The proposed SIM maps acoustic signals captured by the robot to a given geometric map of the environment. The proposed framework can be utilized by any robot that has SLAM capabilities (cameras, range lasers, sonar, etc.) to localize sounding targets.
- A metric is introduced to assess the **reliability** of the proposed framework. Reliability is measured in the context of SSL error and the path length.

II. Related Work

A. Map-Based Navigation

Classical goal-based navigation approaches have been studied extensively in the past. They construct a map of the environment while simultaneously localizing and planning paths to goal locations [3], [11]. For environmental map building and localization, the robot must possess sensors for active perception. These sensors include but not limited to sonar [12], range lasers [13], global positioning systems (GPS) [14], and cameras [11]. Unlike past work, the proposed framework provides acoustic capabilities to mobile robots, thus allowing them to localize sound-emitting targets in complex indoor environments.

B. Audio-Visual Navigation in Unseen Environments

Recent work on the embodied audio-visual task focus on integrating audio and visual cues for navigating through an unseen environment [9], [10], [15], [25]. Gan et al. [9] predict the goal position of the sound source from the audio input and plan an optimal path using raw RGB inputs. Chen et al. [10] learn a policy to predict the best immediate next action using raw audio-visual inputs. Additionally, Chen et al. [15] propose a way-point based technique along with an acoustic memory to enhance the performance of the audio-visual navigation task. Furthermore, Chen et al. [25] introduce semantic audio-visual navigation that utilizes an inferred goal descriptor that integrates both spatial and semantic properties of the target. These studies confirm that audio is a strong navigational signal and can be leveraged by intelligent mobile robot for autonomous navigation.

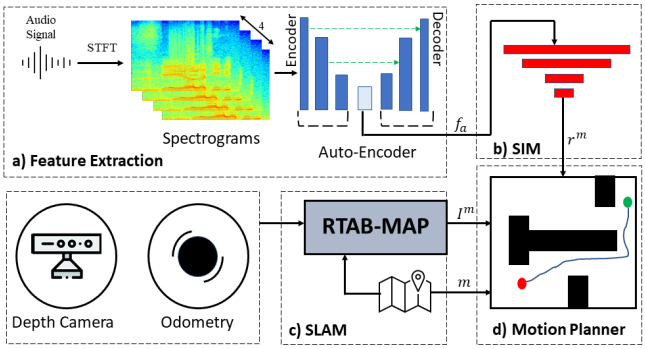


Fig. 2: The proposed framework.

C. Sound Source Localization (SSL)

Past work on SSL focus on different signal processing methods [16], [17], which estimates Direction of Arrival (DOA) and distance to localize the sound source. Common techniques include beamforming, Generalized Cross-Correlation with Phase Transform (GCC-PHAT), Multiple Signal Classification (MUSIC), etc. More recently, deep learning techniques were introduced for DoA estimation [18], [19], [20]. In contrast, the proposed approach in this paper can predict the absolute coordinates of the sound source in a given reference map. There is also considerable work on combining SSL with SLAM, namely Acosutic SLAM (aSLAM) [21]. aSLAM is a technique used to localize the trajectory of a mobile microphone array while simultaneously localizing surrounding acoustic sources. The proposed framework goes beyond that, since the robot must associate the perceived acoustic signal to the constructed geometric map of the environment and plan an optimal path.

III. Audio-Aware Navigation

In this section, we describe the proposed framework (refer Fig. 2) that utilizes DNNs for high resolution SSL. Let $g^m = \begin{bmatrix} g_x^m & g_y^m \end{bmatrix}^\intercal$ be the sound source position and $I^m = \begin{bmatrix} I_x^m & I_y^m & I_\theta^m \end{bmatrix}^\intercal$ denote the robot position w.r.t a given reference map m. Given an acoustic signal captured by an array of 4 microphones, the goal is to localize the acoustic source in m and generate a trajectory from I_m to g_m. Let $x_j(t)$ be the audio signal captured by the jth microphone which is sampled at a frequency f_s. The sampled audio signal is transformed to its amplitude spectrum by computing its Short-Time Fourier Transform (STFT) and is denoted by $X_j \in \mathbb{R}^{F \times T}$, where T is the time dimension and F is the number of frequency bins. Since the spectrum representation of the input signal is of high dimension and contains noise, auto-encoders are used to extract low-dimensional embedding. The feature embeddings f_a extracted from each channel of the microphone array are stacked together, $X_{stacked}$, and fed to the SIM. The goal of the SIM is to learn a DNN that estimates the relative position of the sound source r^m (refer section *3.C*) on a training set $X = (X_{stacked}, r_{gt}^m)$ via supervised learning - $f_\phi(X_{stacked}) = r^m$. The absolute

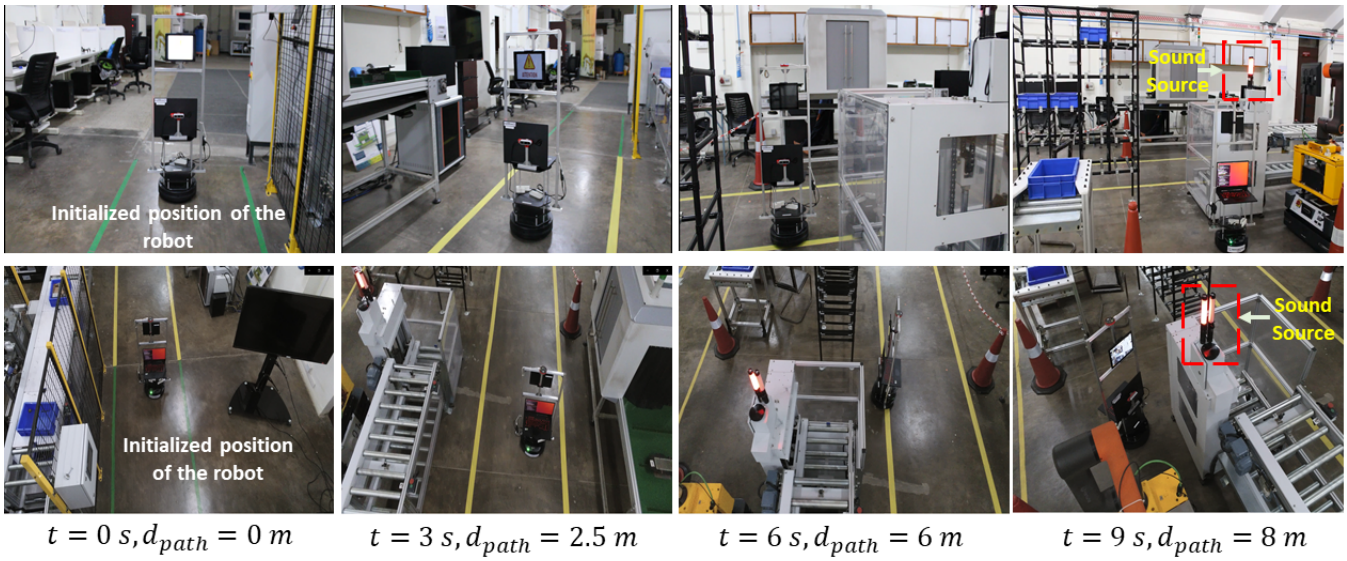


Fig. 3: Illustration of a successful navigation instance. The robot is initialized at a random position ($t = 0s$) and perceives the audio signal. The SIM predicts the relative location of the sound source. The path planner generates a path to the sound source by utilizing the geometric map of the physical environment. The robot successfully navigates towards the sound source ($t = 9s$). (Top) The corresponding third-person views. (Bottom) Key frames from the top view.

position of the source is obtained from the predicted relative location. A path planner generates an optimal path from I^m to g^m. Fig. 3 illustrates an example of a successful navigation instance.

A. RTAB-Map

RTAB-Map [23] is used to construct the 3-D structure of the environment. It utilizes a combination of the robot's odometry and on-board depth camera to simultaneously generate a map, localize in it and plan paths. The current position of the robot I^m is obtained from RTAB-Map.

B. Feature Extraction

To pre-process the raw audio signal sampled at 16000 Hz and a frame size of 1000 ms, the log-scale Short-Time Fourier Transform (STFT) with an FFT size of 1024, hop length of 512 samples and a windowed signal length of 1024 samples is computed and normalized to $[0, 1]$ [22]. The normalized spectrum is resized to a 256×256 complex-valued matrix. While the matrix representation contains some pertinent sound source information, it suffers from the following problems - 1) large dimension size and 2) high Noise-to-Signal ratio. Similar to Liu et al. [22], an auto-encoder is used to address these issues. The auto-encoder is used to extract low-dimensional feature embeddings from the audio spectrogram. The encoder consists of 8 convolutional layers, where each layer is followed by a Leaky-ReLU activation function and batch normalization. The decoder is symmetrical to the encoder structure. The decoder minimizes the Mean Square Error (MSE) between the original input normalized spectrogram and the reconstructed spectrogram. The MSE loss function is defined as -

$$L_{MSE}(s, \hat{s}) = \frac{1}{N}\sum_{i=1}^{N}(s_i - \hat{s_i})^2 \tag{1}$$

where s_i denotes the ith original spectrogram and $\hat{s_i}$ the corresponding reconstructed spectrogram by the decoder. During the reconstruction process, the encoder stores vital information of the audio signal. The implicit features of each channel are extracted as a 256-dimensional embedding f_a. The encoded embedding of all the 4-channels of the microphone array is stacked together and fed to the SIM.

C. Sound Inference Module (SIM)

To regress the relative position $r^m = (x_{relative}, y_{relative})$ of the sound source, we train the SIM on Euclidean loss. r^m indicates that the sound-emitting target is located at a distance of $x_{relative}$ meters to the right of the robot and $y_{relative}$ meters in front of the robot [9]. The stacked feature embeddings of the audio spectrograms are fed to the SIM for training. The embeddings are passed through a series of fully-connected layers. The detailed architecture of the SIM consists of FC(256)-FC(128)-FC(64)-FC(2). Each layer consists of a ReLU activation function except for the last layer. The last layer consists of a tanh activation function and the network is trained using the MSE loss function. The MSE loss is calculated between the predicted relative location and the ground truth relative position of the sound source with reference to the given geometric map. The optimizer used is ADAM [27], with a learning rate of 1×10^{-3} and a batch size of 64. For better loss convergence, the ground truth relative positions are normalized to $[-1, 1]$. The auto-encoder and the SIM are trained separately.

D. Motion Planner

The robot utilizes the ROS navigation stack for global path planning. Dijkstra's algorithm is implemented to generate a global path from the robot's current position to the absolute position predicted by the SIM. A topological map representation is used to define goal locations and plan collision free trajectories [23]. This representation is denoted by a tuple $T =< N, E, A, nav >$, where $N = [n_1, n_2, n_3, ..., n_k]$ is defined as a set of k nodes that represent positions in the physical space, $E \subseteq N \times N$ denotes the set of all edges that connects the nodes, A is the set of all actions that can be executed by the robot, and $nav : E \rightarrow A$ associates each edge to a navigation action.

IV. Experiment

A. Setup

The proposed methodology is validated on a Turtlebot 2 platform (refer Fig. 4). A circular 4-channel ReSpeaker Mic Array[1] is mounted at the height of $1.2m$ from the robot's base. The diameter of the microphone array is $7cm$. The robot is also equipped with an Intel Realsense D435i depth camera for SLAM. The Turtlebot 2 is connected to an on-board laptop that runs online in ROS. All experiments were conducted in an experimental industrial testbed to verify the robustness of the presented approach. In order to test if the framework can generalize to different map configurations, the original physical environment was subjected to minor modifications such as a change in the number of personnel and/or the addition of some bulky objects. Additionally, two other scenarios were considered - 1) sound source recorded without background noise, and 2) sound source recorded in the presence of different real-world distractor sounds commonly found in factories. The SIM was trained in three and tested in two different environmental configurations. 10 random locations and orientations (at a distance of 1m, 2m,...,5m) were selected during testing. The same map was used during evaluation. 3 random sound source positions were selected that are not used during training.

B. Data Collection

Training data for the SIM was collected in the same smart factory setup . The given geometric map is discretized into grids for easy quantitative evaluation [5]. For training, four different static sound source positions were chosen (two with and two without background noise). For each sound source position, four array orientations (0°, 90°, 180°, and 270°) and 25 traversable grid locations are selected. At each position, the robot records the perceived audio signal from the on-board microphone array, and the relative position between the robot's current position and the sound source is noted. The entire training data generation process is illustrated by Algorithm 1. Since collecting a vast amount of audio signals is cumbersome, we incorporate a data augmentation technique. At each data collection point, audio signals are recorded for T seconds and smaller patches

[1]https://www.seeedstudio.com/ReSpeaker-Mic-Array-v2-0.html

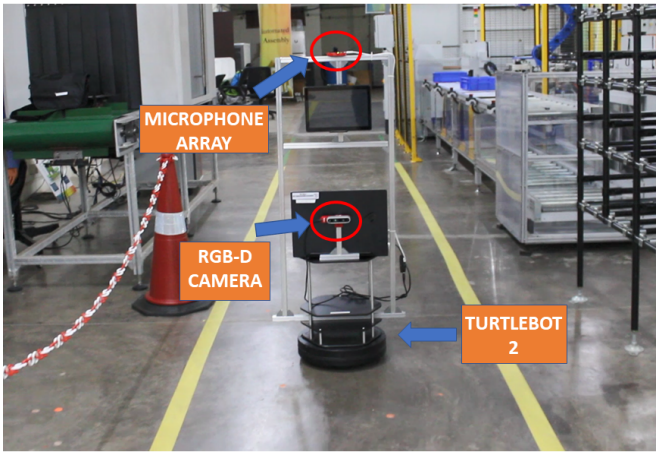


Fig. 4: Experimental setup for the audio-aware navigation task. A Turtlebot 2 is equipped with a 4-channel microphone array and an Intel Realsense D435i RGB-D camera.

Algorithm 1: Generate Training Data

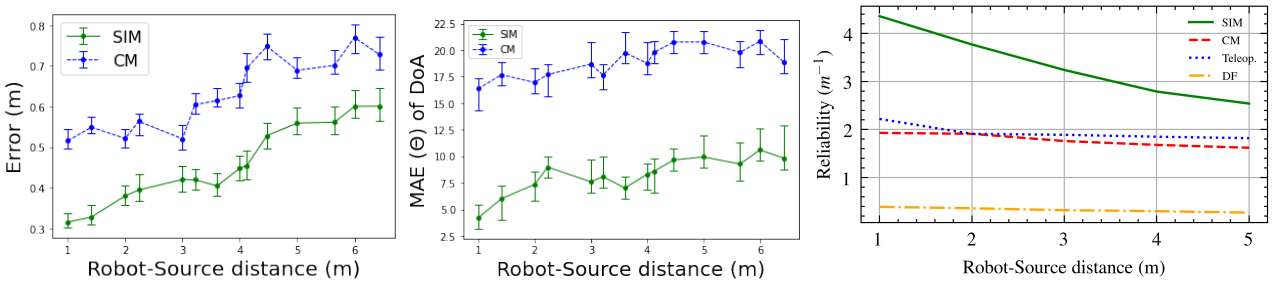
Require: Microphone array with M microphones and a geometric map of the environment

Discretize the map into grid worlds
Select N random sound source positions
for each $n \in N$ **do**
  Select P array orientations
  Select G traversable grid locations
  for each $p \in P$ **do**
    for each $g \in G$ **do**
      Record audio signal for M mics
      Calculate the relative distance between source and array
      Save: Audio Recording
    end for
  end for
end for

of one second $x_1, x_2, x_3, ..., x_n$ of the underlying audio signal are extracted in a sliding window fashion where $x_n \in \mathbb{R}^t, t < T$. Since the factory setup is subjected to background noise, slicing the signal into smaller patches could help to generalize over varying background noise. All the experiments were conducted in a $125m^2$ physical world with a smart factory setup. The environment is highly complex and consists of several manufacturing machinery and industrial robots. Trials are conducted with a stationary sound source located at the height of $1.25m$. The sound source is a loudspeaker speaker emulating different sounds commonly found in a factory.

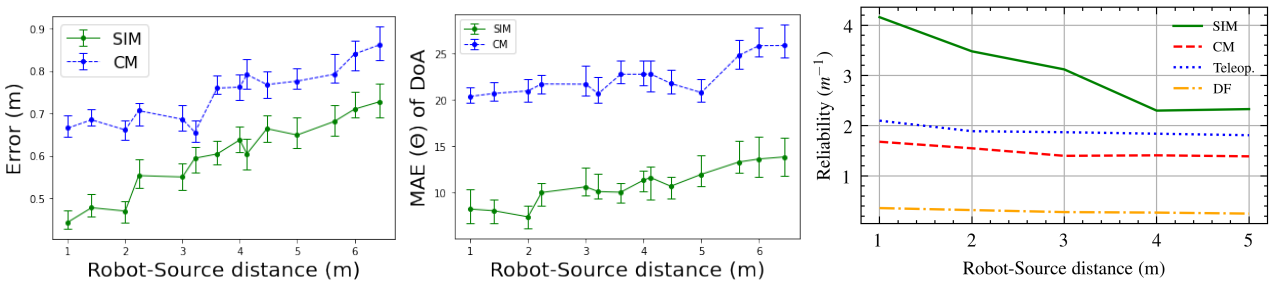
C. Evaluation Metrics

1) Error (e_d): This metric is used to evaluate the localization performance of the proposed SIM at a given distance d from the sound source. (e_d) is defined as follows for N test episodes -



(a) Error over Robot-Source distance (b) MAE over Robot-Source distance (c) Reliability over Robot-Source distance

Fig. 5: Trials in test scene without background noise



(a) Error over Robot-Source distance (b) MAE over Robot-Source distance (c) Reliability over Robot-Source distance

Fig. 6: Trials in test scene with background noise

$$e_d = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(x_{gt} - x_{pred_i})^2 + (y_{gt} - y_{pred_i})^2} \quad (2)$$

where, (x_{gt}, y_{gt}) is the ground truth position of the sound source and (x_{pred_i}, y_{predi}) is the absolute location of the sound source. Since the SIM outputs the relative coordinates of the sound source with respect to the robot's current position, the Direction of Arrival (DoA), θ_d can also be obtained by simple calculations. The Mean Absolute Error (MAE) between the predicted DoA and the ground truth orientation is considered for evaluation.

2) Reliability (r_d): It is a metric that reflects the reliability of the proposed framework. To define such a measure, a reliability criterion is adopted. Similar to [24], the metric evaluates the navigation performance of the model. The reliability metric for N test episodes is defined as follows -

$$r_d = \frac{1}{N}\sum_{i=1}^{N}\frac{d_{shortest}}{e_d \times \max(d_{shortest}, d_{path_i})} \quad (3)$$

The metric penalizes the model for large errors and weighs successful navigation instances with the ratio of the shortest path $d_{shortest}$ to the executed path d_{path}.

D. Baseline

Three baselines methods were considered for validation -

1) Teleoperation: The robot is teleoperated from a distance by an operator. The operator relies on the Audio-Visual feedback from on-board sensors for navigation.
2) Direction Follower (DF) [10]: The robot sets intermediate goals K meters away in direction of the perceived signal. The navigation instance for this baseline is considered completed if it comes within 2 cm of an obstacle or if the navigation time exceeds a certain limit.
3) Convolutional Mapper (CM) : In this setting, the proposed SIM is replaced with Gan et al's [9] sound perception module which implements a VGG like Convolutional Neural Network (CNN).

V. Results

Table 2 summarizes the reliability in different scene configurations. The proposed framework outperforms all baselines in both training and testing scenarios (see Fig. 5 (c) and 6 (c)). During testing, our model achieves over a 58% reliability on average for both acoustic conditions, i.e with and without background noise. It is not surprising to find that the reliability of the DF baseline is very low because it was unable to anticipate the position of the sound source. The teleoperation baseline is better than DF, as the operator can anticipate the position of the sound source through Audio-Visual feedback from on-board sensors for navigation. However, the performance of this baseline is relative and depends solely on the operator. The CM baseline is comparable to teleoperation in terms of reliability;

TABLE I: Training time for 400 epochs, average accuracy over a distance of 6m, and processing time.

	SIM	CM
Training time	3.5 min	2.5 min
Accuracy	77.8%	64.4%
Processing time (per batch)	46 ms	38 ms

however, it performs poorly at large distances. In contrast, the proposed framework utilizes auto-encoders that extract implicit features from the audio signal robust against a high noise-to-signal ratio, ad thus accurately localizing the source. Additionally, to understand the robustness of the proposed framework under microphone noise, reliability is validated for different noise levels at a Robot-Source distance of 6 m (refer Fig. 8). For microphone noise, an increasing Gaussian noise is added to the perceived audio signal. The proposed approach outperforms all other baselines and is robust in noisy environments. Table 1 reports the runtime and average positional accuracy of prediction for the two DNN based methods. Both the methods were trained for 400 epochs. Processing time is an important factor to consider if we want to implement a DNN-based auditory navigation system in real-time. The average time for the SIM and CM is about 46 ms and 38 ms, respectively. Both the models are tested on a machine with 3.3GHz AMD Ryzen 9 5900HS and an NVIDIA GeForce RTX 3060. While the SIM is slower than the CM, the accuracy is higher.

Figures 5 (a,b) and 6 (a,b) illustrate the average localization performance observed with respect to the distance between the robot and source, incrementing 1 meter across all testing scenarios. Error (e_d) and Mean Absolute Error (MAE) in the DoA are validated for the SIM and the CM baseline. It is observed that the SIM performs better than the CM baseline in both acoustic conditions (Error is 29.46% and 24.56% lower, with and without background noise, respectively). Results indicate that the proposed SIM can accurately predict the location of the sound source with an error less than 1 meter and an MAE in the predicted DoA less than 15 degrees for distance within 7 meters. Furthermore, it is observed that the error and MAE in DoA increase with an increment in the robot-source distance. As the distance between the robot and the source increases, the complexity of the physical environment increases. Consequently, the perceived signal suffers from multiple reflections and is more susceptible to environmental noise and reverberation. Therefore, both prediction models are prone to a higher value of error. However, the observed error in the SIM remains below 1 meter.

VI. CONCLUSION

In conclusion, this paper presents the audio-aware navigation task of a mobile robot in a complex indoor environment. A mobile robot equipped with a microphone array can utilize the framework to efficiently navigate to targets in previously unseen environmental configurations of the same scene. We demonstrate the potential of DNNs for sound source localization in a real-world environment. Deep learning methods can learn characteristic reverberations for every position to separate the signal. Experimental results indicate that the proposed framework localizes sound sources with an average positional error of less than 0.55m, which significantly outperforms the three baselines.

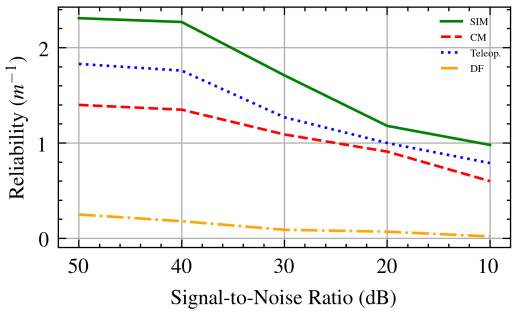


Fig. 7: Reliability vs Microphone Noise at a Robot-Source distance of 6 m.

TABLE II: An illustration of reliability for different audio-aware navigation methods.

Baseline	d_{r-s}	Map Config. (train)		Map Config. (test-1)		Map Config. (test-2)	
		W/N	N	W/N	N	W/N	N
Teleoperation	1.0	2.29	2.14	2.20	2.10	2.23	2.10
	2.0	1.94	1.87	1.90	1.85	1.91	1.89
	3.0	1.90	1.86	1.87	1.86	1.89	1.87
	4.0	1.89	1.88	1.89	1.87	1.86	1.85
	5.0	1.87	1.85	1.86	1.82	1.81	1.83
Direction Follower (DF)	1.0	0.43	0.32	0.38	0.33	0.39	0.36
	2.0	0.38	0.31	0.37	0.30	0.34	0.33
	3.0	0.32	0.30	0.32	0.33	0.32	0.32
	4.0	0.30	0.29	0.29	0.27	0.31	0.28
	5.0	0.28	0.27	0.26	0.25	0.27	0.25
Conv. Mapper (CM)	1.0	3.42	3.16	1.90	1.56	1.95	1.69
	2.0	3.35	3.05	1.82	1.68	1.90	1.57
	3.0	3.12	2.94	1.56	1.30	1.86	1.31
	4.0	2.88	2.73	1.59	1.36	1.78	1.42
	5.0	2.68	2.63	1.58	1.45	1.62	1.40
Sound Inference Module (SIM)	**1.0**	**4.76**	**4.24**	**3.49**	**4.19**	**4.36**	**4.15**
	2.0	**4.49**	**4.36**	**3.91**	**3.84**	**3.77**	**4.02**
	3.0	**3.92**	**3.82**	**3.21**	**2.95**	**3.24**	**3.48**
	4.0	**3.85**	**3.55**	**2.39**	**2.25**	**2.79**	**2.48**
	5.0	**2.98**	**2.64**	**2.47**	**2.15**	**2.54**	**2.31**

VII. ACKNOWLEDGEMENT

We acknowledge the support from the Common Engineering Facility Centre (CEFC) at the Centre for Product Design and Manufacturing, Indian Institute of Science Bangalore and the Department of Heavy Industries (DHI), Govt. of India.

REFERENCES

[1] Tolimieri, Nick, Andrew Jeffs, and John C. Montgomery. "Ambient sound as a cue for navigation by the pelagic larvae of reef fishes." Marine Ecology Progress Series 207 (2000): 219-224.

[2] Merabet, L.B., Pascual-Leone, A.: Neural reorganization following sensory loss: the opportunity of change. Nat. Rev. Neurosci. 11, 44–52 (2010).

[3] Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics. MIT Press, Cambridge (2005).

[4] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu et al., "Learning to navigate in complex environments," in ICLR, 2017.

[5] 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 ICRA, 2017.

[6] Gupta, S., Davidson, J., Levine, S., Sukthankar, R., Malik, J.: Cognitive mapping and planning for visual navigation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2616–2625 (2017).

[7] Savva, M., et al.: Habitat: a platform for embodied AI research. In: ICCV (2019).

[8] Mishkin, D., Dosovitskiy, A., Koltun, V.: Benchmarking classic and learned navigation in complex 3D environments. arXiv preprint arXiv:1901.10915 (2019).

[9] Gan, Chuang, et al. "Look, listen, and act: Towards audio-visual embodied navigation." 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020.

[10] Chen, Changan, et al. "Soundspaces: Audio-visual navigation in 3d environments." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. Springer International Publishing, 2020.

[11] Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cam- bridge University Press, Cambridge (2004).

[12] Ribas D, Ridao P, Tardós JD et al (2008) Underwater SLAM in manmade structured environments. J Field Robot 25(11):898–921.

[13] Thrun S, Montemerlo M, Aron A (2006) Probabilistic terrain analysis for high-speed desert driving. In: Proceedings of robotics: science and systems.

[14] Thrun S, Montemerlo M, Dahlkamp H et al (2005a) Stanley: the robot that won the DARPA grand challenge. J Field Robot 23(9):661–692.

[15] Chen, Changan, et al. "Learning to set waypoints for audio-visual navigation." arXiv preprint arXiv:2008.09622 (2020).

[16] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[17] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," Robotics and Autonomous Systems, vol. 96, pp. 184–210, 2017.

[18] P. Pertil¨a and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

[19] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," Journal of Robotics and Mechatronics, vol. 29, no. 1, pp. 37–48, 2017.

[20] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

[21] C. Evers and P. A. Naylor, "Acoustic SLAM," IEEE/ACM Trans. Audio, Speech, Language Proces., vol. 26, no. 9, pp. 1484–98, Sep. 2018.

[22] H. Liu, Z. Zhang, Y. Zhu and S. -C. Zhu, "Self-Supervised Incremental Learning for Sound Source Localization in Complex Indoor Environment," 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 2599-2605, doi: 10.1109/ICRA.2019.8794231.

[23] J. P. Fentanes, B. Lacerda, T. Krajnik, N. Hawes, and M. Hanheide, "Now or later? Predicting and maximising success of navigation actions from long-term experience," in Proc. - IEEE Int. Conf. Robot. Autom., vol. 2015-June, no. June, 2015, pp. 1112–1117.

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

[25] Chen, Changan, Ziad Al-Halah, and Kristen Grauman. "Semantic Audio-Visual Navigation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[26] Purushwalkam, Senthil, et al. "Audio-visual floorplan reconstruction." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[27] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.