2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Sim2Real Transfer for Audio-Visual Navigation with Frequency-Adaptive Acoustic Field Prediction

Changan Chen*, Jordi Ramos*, Anshul Tomar*, Kristen Grauman
University of Texas at Austin

***Abstract*— Sim2real transfer has received increasing attention lately due to its success in transferring robotic policies learned in simulation to the real world. While significant progress has been made in transferring vision-based navigation policies, the current sim2real strategy for audio-visual navigation remains limited to basic data augmentation. Sound differs from light in that it spans across much wider frequencies and thus requires a different solution for sim2real. To understand how the acoustic sim2real gap varies with frequencies, we first define a novel acoustic field prediction (AFP) task that predicts the local sound pressure field. We then train frequency-specific AFP models in simulation and measure the prediction errors on collected real data. We propose a frequency-adaptive strategy that intelligently selects the best frequency band for prediction based on both the measured prior and the energy distribution of the received audio, which improves the generalization on real data. Coupled with waypoint navigation, we show the navigation policy not only improves navigation performance in simulation but also transfers successfully to real robots. This work demonstrates the potential of building autonomous agents that can see, hear, and act entirely from simulation, and transferring them to the real world.**

I. Introduction

Navigation is an essential ability of autonomous robots, allowing them to move around in the environment and execute tasks such as delivery, search and rescue. Sometimes, the robot also needs to hear the environment and navigate to find where the sound comes from, e.g., when someone is asking for help in a house, or when the fire alarm goes off.

Navigation has been extensively studied in the robotics community and has been traditionally approached with Simultaneous Localization and Mapping (SLAM) [1] with Lidar sensors, which is limited in semantic reasoning. Recent research has increasingly focused on vision-centric navigation, where robots rely primarily on visual sensors to perceive their environment, showing significant success in photorealistic real-scanned settings [2]. Various tasks have been proposed, such as PointGoal navigation [3], [4], [5], ObjectGoal navigation [6], [7], or visual exploration [8], [9], [10]. Other work further expands the sensory suite to include hearing. In particular, the AudioGoal task [11], [12] requires an agent to navigate to a sounding target (e.g., a ringing phone) using audio for directional and distance cues while using vision to avoid obstacles in the unmapped environment.

With the success of these learning-based navigation systems in simulation, efforts have been made to transfer learned policies to the real world by addressing the sim2real gap [13], [14], [15]. However, recent work [16] on audio-visual navigation with sim2real transfer through data augmentation, which fails to account for the influence the range of frequencies have on the sim2real gap. In this work, we systematically evaluate the acoustic gap and propose a solution to bridge it.

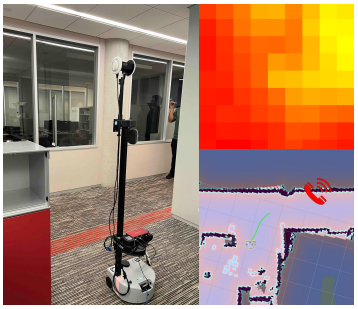

Fig. 1: Our robot predicts an acoustic field with a frequency-adaptive model and navigates to locate the sound source.

State-of-the-art approaches in audio-visual navigation rely on reinforcement learning to train the navigation policy end-to-end [11], [17], which not only poses challenges for interpretation but also struggles to generalize to the real world due to various sim2real gaps. Recent advancements in visual navigation have shown successful sim2real transfer with hierarchical models [13], [8], which consist of a high-level path planner and a low-level motion planner. This hierarchical design helps mitigate some of the low-level physical discrepancies encountered during transfer. However, existing hierarchical models have not yet attempted to address sim2real for audio-visual navigation.

Inspired by these methods, we design a modular approach to ease the transfer from simulation to the real world. To achieve this, we confront a key question: what is the proper high-level planning task that can survive sim2real transfer for audio-visual navigation? To this end, we propose a novel prediction task: *acoustic field prediction*—predicting the local sound pressure field around the agent. The gradient of this field reflects the direction of the sound. Measuring acoustic fields is expensive in the real world since it requires simultaneously capturing the sound pressure of all points in the field due to the dynamic nature of sound. However, computing acoustic fields is free in simulation. We first build an audio-visual model as the acoustic field predictor (AFP) and curate a large-scale acoustic field dataset on SoundSpaces 2.0 [18], the state-of-the-art audio-visual simulation platform. We show that this approach outperforms existing methods on the Continuous AudioGoal navigation benchmark.

After validating the proposed approach in simulation, we

* indicates equal contribution, sorted in alphabetical order

then investigate where acoustic discrepancy arises. It is known that ray-tracing-based acoustic simulation algorithms introduce more errors with lower frequencies due to wave effects [19]. Given this observation, we focus on evaluating how sim2real errors vary with frequencies. We first collect real acoustic field data with the source sound being white noise, whose audio energy uniformly spans across all frequencies. We then train acoustic field prediction models that only take the sub-frequency band of the input audio and test it on the real white noise data. By computing the errors across multiple samples, we show that the errors do not strictly go down as the frequency goes higher, and using the best frequency band yields errors smaller than using all frequencies for the white noise sound.

However, simply taking the best frequency band does not work for all sounds since different sounds have varying spectral distributions. To address this issue and make the model aware of the spectral difference, we propose a novel frequency-adaptive prediction strategy, which selects the optimal frequency sub-band based on measured errors and the spectral distribution of the received audio. To validate this approach, we collect more acoustic field data with various sounds and show that the frequency-adaptive model leads to the lowest error on the real data compared to other strategies.

Lastly, we build a robot platform that equips the Hello Robot with a 3Dio binaural microphone and then deploy our trained policy on this robot. We show that our robot can successfully navigate to various sounds with our trained frequency-adaptive acoustic field prediction model. See Fig. 1 and Supp. video.

In summary, we propose a novel acoustic field prediction approach that learns to navigate without interaction with the environment. This approach improves the SOTA methods on the challenging Continuous AudioGoal navigation benchmark [18]. We perform a systematic evaluation of the sim2real challenge and propose a frequency-adaptive strategy as the treatment for sim2real. We show this strategy works on both collected real data as well as on our robot platform. To the best of our knowledge, this is the first work to investigate and propose a principled solution to the sim2real transfer problem for audio-visual navigation.

II. Related Work

A. Embodied Navigation

To navigate autonomously, traditionally a robot builds a map via 3D reconstruction (i.e., SLAM) and then plans a path using the map [20]. Recent works have developed navigation policies that make navigation decisions in a previously unmapped environment from egocentric observations directly without relying on mapping [4], [7], [21]. Some recent efforts have developed audio-visual simulation platforms [11], [18], [22] that enable embodied agents to both see and hear. In the AudioGoal navigation task [11], [17], the agent must navigate to the source of sound in an unknown environment. The state-of-the-art audio-visual navigation models train policies with reinforcement learning (RL), requiring millions of samples. Inspired by recent work in learning potential functions for interaction-free navigation [8], we propose to predict acoustic fields for interaction-free audio-visual navigation.

B. Sound Localization in Robotics

In robotics, microphone arrays are often used for sound source localization [23], [24], [25]. Past studies fuse audio-visual cues for surveillance [26], [27], speech recognition [28], human robot interaction [29], [30], and robotic manipulation tasks [31]. These solutions typically depend on analytical solutions for computing the direction of sound, whose performance deteriorates under strong reverberation and noise [32]. Recent works propose learning-based sound localization [33], [34], which however require collecting real data for training. In this paper, we show that we can transfer models trained in simulation directly to the real world.

C. Sim2real Transfer

Benefiting from recent large-scale datasets of real-world 3D scans [35], [36] and supporting simulators [5], [37], recent works have shown success in enabling embodied agents in simulation. Transferring the model trained in simulation to the real world is thus of great interest. The mostly widely used approaches are domain randomization [38], [39], system identification [15], [40], and transfer learning and domain adaptation [41]. Most sim2real research studies transferring a policy from simulation to the real world based on visual input. Recent work [16] does sim2real transfer for audio-visual navigation by applying data augmentation empirically, which does not account for the effect of frequencies on sim2real. In this work, we systematically evaluate the sim2real gap by collecting data and identifying the spectral discrepancy.

III. Approach

A. SoundSpaces Platform and Audio-Visual Navigation

We first introduce the simulator that we are using. SoundSpaces 2.0 [18] is a state-of-the-art audio-visual simulation platform that produces highly realistic audio and visual rendering for arbitrary camera and microphone locations in 3D scans of real-world environments [36], [42], [35]. It accounts for all major real-world acoustics phenomena: direct sounds, early specular/diffuse reflections, reverberation, spatialization, and materials and air absorption.

In audio-visual navigation, the goal is to navigate to a sounding object in an unknown environment by seeing and hearing. The location of the sound source is not known and needs to be inferred from the received audio. At every step, the agent needs to sample an action from {MOVE_FORWARD, TURN_LEFT, TURN_RIGHT, STOP}. If the agent issues the STOP action within a 1m radius of the goal, the episode is considered successful.

There are different instantiations of the audio-visual navigation task, each with its goal specification. For example, AudioGoal [11] requires navigating to a static target in a discretized environment, while Dynamic AudioGoal [43] requires navigating to a moving sound source. We target the Continuous AudioGoal navigation benchmark introduced in SoundSpaces 2.0 [18], which generalizes the state space to be the continuous environment.

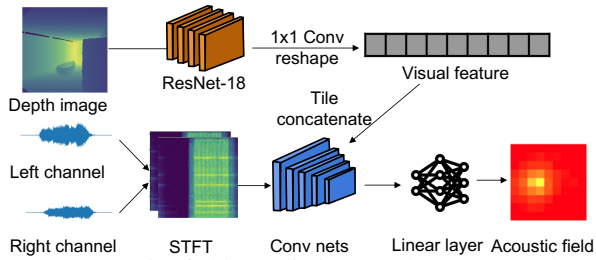

Fig. 2: Acoustic field prediction model. The model first extracts audio and visual features, and then tiles and concatenates features channel-wise to predict the acoustic field.

B. A Modular Design for Sim2real Transfer

Transferring a navigation policy trained in simulation to the real world is not trivial due to many domain gaps between the simulation environment and the real world, which include the visual discrepancy, the physical dynamics discrepancy, the robot actuation discrepancy and—specifically in this task—the acoustic discrepancy.

We focus on investigating the acoustic discrepancy and to bridge other domain gaps (e.g., visuals and physics), we take a hierarchical approach that disentangles navigation into high-level path planning and low-level motion planning. This has a few benefits: 1) disentangling the policy makes it possible to utilize existing SLAM algorithms on the real robot to abstract away domain gaps other than the audio. 2) disentangling the policy makes the intermediate output more interpretable and easier to debug 3) specifically in this work, posing the high-level planning as a supervised prediction task makes it easier to measure the sim2real difference because we can evaluate the performance by collecting real measurements without repeatedly running robots.

The key challenge here is to formulate the proper waypoint prediction task that could survive the sim2real transfer. One existing approach [12] predicts the exact location of the audio goal directly, which is however an ill-posed problem since the environment geometry is unknown. For example, when the audio goal is in another room, the received audio reveals the direction to the door rather than the exact direction of the goal. Instead of predicting the global audio goal location, we propose to predict the local acoustic field (sound pressure field) centered around the agent, which not only better captures the direction of the sound but also is more predictable from the visual observation of the environment.

Defining the waypoint prediction task however does not address the audio discrepancy directly. SoundSpaces 2.0 renders the audio propagation as a function of the geometry of the environment, the material properties, and source/receiver locations based on a bidirectional ray-tracing algorithm [44]. While it produces realistic audio renderings, there remains some difference between how sound propagates in the real world vs in simulation. It is known that ray-tracing-based algorithms yield worse performance with lower frequencies due to wave effects [19]. This implies the model needs to be aware of this spectral difference for sim2real. Thus we introduce a frequency-adaptive prediction strategy to help the model better transfer to the real world in Sec. III-E.

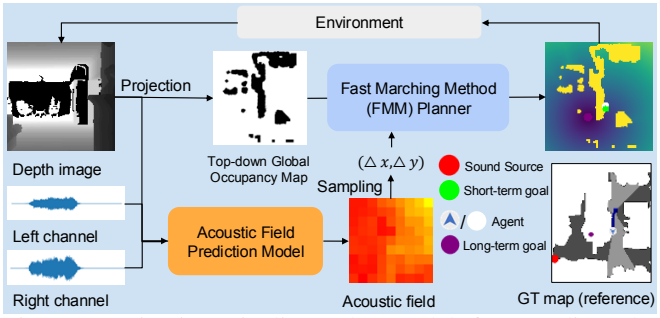

Fig. 3: Navigation pipeline. The model first predicts the acoustic field, samples the peak as the long-term goal, and navigates toward the goal with a path planner.

C. Acoustic Field Prediction

We define acoustic field prediction as predicting the top-down sound pressure field of $L \times L$ centered at the agent, given the egocentric depth image of 128×128 pixels and a one-second binaural audio.

To tackle the acoustic field prediction problem, we first present a model that uses both audio and visual observations (see Fig. 2). The motivation behind using the visual sensor is that the visual observation of the surrounding environment can be useful in inferring the geometry of the environment, which affects the acoustic field. For example, walls often act as the boundary of the acoustic field.

More specifically, we use a pre-trained ResNet [45] to extract features from the input image and reshape it with a 1×1 Conv layer into a 1d vector of size 512. For the input binaural audio, we first process the waveforms with Short-Time Fourier Transform (STFT) to convert the time-domain signal into the frequency domain. We then use a 2D Conv net to encode the features and then tile and concatenate with the visual feature. Lastly, we feed the final output through one linear layer and reshape the prediction into the size of the target acoustic field $L \times L$.

D. Hierarchical Navigation

With this acoustic field prediction model, we then construct a hierarchical navigation pipeline (see Fig. 3) to perform audio-visual navigation, which executes the following steps: 1) sampling a long-term goal; 2) navigating to the long-term goal; and 3) making the stopping decision.

1) Sampling a Long-Term Goal: At each time step, the agent predicts the acoustic field based on audio-visual inputs and then identifies the maximum value of the field. We set the peak location as the long-term goal either when there is no existing long-term goal or the new peak value surpasses the value of the existing long-term goal since as the agent gets closer to the goal, the sound usually gets louder.

2) Navigating to the Long-Term Goal: After sampling the long-term goal, for path planning, we use the Fast Marching Method (FMM) [46] to determine the best route to the goal in simulation. FMM takes the occupancy map that is built on the fly, the agent's current position, and the long-term goal as inputs. The occupancy map is computed by calculating the point cloud observed at each timestep using the depth camera. Next, FMM calculates the distance between each navigable point in the map to the long-term goal. The

algorithm then selects the adjacent point on the map with the lowest value as its short-term goal and the agent then moves towards that point. When the long-term goal is sampled at a non-navigable location, we use breadth-first search (BFS) to find the closest available point to navigate to.

3) Stopping Criteria: The stopping condition is evaluated each time after the agent reaches a long-term goal or the closest navigable point to the long-term goal. When the agent samples a new long-term goal, if the peak value of the predicted acoustic field is at the center of the field, the agent issues the stop action.

E. Frequency-adaptive Prediction

Existing audio-visual navigation models use all frequencies in the input audio. However, sound spans across a wide range of frequencies, and due to imperfect geometric simulation techniques, the acoustic gap varies as a function of frequencies. Models trained with all frequencies assuming them equally reliable would have lower performance when deployed on a real robot.

Given this observation, we first systematically examine how the gap changes as a function of the frequency. The idea is simple: with a given frequency band $[F_1, F_2]$, we first train an acoustic field prediction (AFP) model in simulation using only that band, then test it on real-world data of the same sound and same band, and calculate the prediction error. We equally divide all frequencies into N subbands ($N = 5$ based on hyperparameter tuning), and we show the distribution of errors over the frequency bands in Fig. 4, where error is defined as the distance between the predicted max and ground truth goal location. As expected, the lower frequencies tend to yield larger prediction errors. However, the error does not monotonically decrease as frequency increases likely due to simulation errors. We also trained a model that uses all frequencies, which has a distance error of $0.86m$, underperforming the best frequency band.

With this measurement, the most intuitive idea would be just to take the frequency band that has the least sim2real error and train a model with that band. However, this will not work for real-world scenarios where some sounds span across many frequency bands while others only occupy a very narrow range of frequencies. To take that into account, we propose a frequency-adaptive prediction strategy that uses the best frequency band based on both the measured error and the energy distribution of the received audio.

Assume we divide all frequencies linearly into N bands. Given a received audio A_r, we first convert it into the frequency domain and divide it into these N bands. Based on the measured errors, we have a weighting function that assigns weights to these bands based on their sim2real errors:

$$p(i) = (\frac{1}{e_i})^{\alpha}, i \in [1, ..., N], \quad (1)$$

where e_i is the error in Fig. 4. For each subband i of the input, we then compute another weight based on the energy of the band normalized with respect to the highest energy:

$$q(i) = (\frac{r_i}{r_m})^{\beta}, i \in [1, ..., N], \quad (2)$$

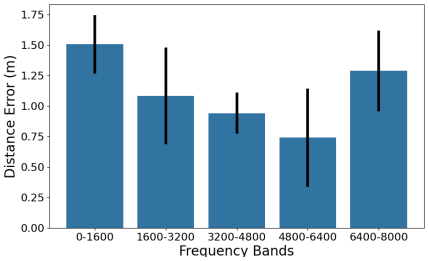

Fig. 4: Sim2real error as a function of frequencies. We report the mean and standard deviation of distance errors between the predicted and the ground truth peak locations.

where r_i is the energy of that band and $r_m = \max_i r_i$. We basically assign higher weights to frequency bands that have more energy. Lastly, we take the product of these weights:

$$w(i) = p(i) \times q(i), i \in [1, ..., N] \quad (3)$$

We take band i with the highest $w(i)$ to produce the final prediction. Both α and β are hyper-parameters, and we perform a grid search to find the best values on validation.

Intuitively, what the weighting function does in Eq. (3) is: if the input sound has a fairly equal distribution of energies over all subbands, it will take the best band from $p(i)$ that has the lowest sim2real error. If the input sound has a very skewed energy distribution, it will prioritize taking the band where the audio has the most energy. In this way, we factor into both the measured difference and the spectral distribution of individual sounds.

F. Implementation and Training Details

For the size of the acoustic field, we set L to 9 with a grid resolution of $0.5m$, i.e., $4.5m \times 4.5m$ centered around the agent based on our ablations. α and β are set to 5 and 0.8 respectively based on the validation performance.

We train the predictor with Mean Squared Error (MSE) loss till convergence. For optimization, we use the Adam optimizer [47] with a learning rate set of 0.001.

IV. Data Curation

Due to the expense of measuring real acoustic field data, we choose to utilize simulation to collect large-scale training data. We also collect real data for measuring the sim2real gap and validating our frequency-adaptive prediction model.

A. SoundSpaces Acoustic Field Dataset

SoundSpaces 2.0 supports computing the impulse response $I(s, r)$ between the source location s and the receiver location r as a function of the 3D environment but does not have direct API support for rendering the acoustic field. To compute the field, given a (s, r) pair, we first sample a grid centered at the receiver location of size $L \times L$, and for each grid point p, we compute $I(s, p)$ which results in L^2 number of RIRs per receiver location. However, these RIRs are represented in the form of waveforms instead of single numbers. To best represent the sound pressure at each single point, we take the maximum amplitude of the waveform.

For sampling the source/receiver locations and environments, we utilize the existing audio-visual navigation episodes [11], which provides configurations of the environment and source/receiver locations. This dataset uses scenes

from the Matterport3D dataset [36], which contain scans of real-world environments such as apartments, offices, and even churches. We sample 500 episodes per environment for the 57 training environments in the navigation dataset. We also perform a similar operation to curate the validation and test set. In total, we collect 1.1M/52K/52K samples for train/val/test. Along with these acoustic fields, we also render the RGB-D images at the corresponding locations. See examples in Fig. 6.

B. Real Measurements Collection

To measure the sim2real error, we collect real audio measurements to evaluate the trained model's performance. For that, we use a 3Dio microphone to capture the binaural audio with a smartphone serving as the speaker output. We align the real-world parameters closely with those in our simulator, such as the height of the speaker and receiver. Since the simulator employs a mono receiver, the two-channel audio data we gather is transformed into mono format by averaging the amplitude values across both channels. This process is repeated for ten distinct speaker positions (8 different directions w.r.t the agent and two data points for when the speaker is near the agent). We also downsample the acoustic field resolution from 9×9 to 3×3 so that we can collect more data in more environments.

For the source of the sounds, we use two types of sounds: white noise and normal sounds. To compute the sim2real errors in Fig. 4, it is important for the sound to have uniform distribution across all frequencies, and we use white noise for that. For evaluating the final frequency-adaptive acoustic-field prediction model, we choose 7 unheard sounds that have varying spectral distributions and play them as the source. For each sound, we collect 10 data points. We split them equally into validation and test for hyperparameter searching.

V. Robot Platform

To deploy our sim2real policy on a real robot, we build our audio-visual robot by equipping a HelloRobot with a 3Dio binaural microphone as shown in Fig. 1. We use Focusrite Scarlett Solo as the audio interface to amplify the audio signals from the binaural microphone.

To start the navigation, we first sample the current audio from the microphone and predict the long-term goal from the acoustic field. We then pass this goal to the robot and use HelloRobot's navigation stack to move the robot towards the goal. Once the robot reaches the long-term goal, it comes to a complete stop for a second to sample the audio again. This process is repeated until the predicted goal location is in the center of the acoustic field. If the sampled long-term goal is in an inaccessible region, we have a time limit of 5 seconds after which the robot stops and samples a new goal.

VI. Experiments

A. Results on Continuous AudioGoal Navigation Benchmark

We first demonstrate the effectiveness of our navigation system on the challenging Continuous AudioGoal navigation benchmark [18], where the agent moves in a continuous unseen environment to find the location of a ringing telephone sound. For metrics, we use the common Success Rate (SR), success weighted by inverse path length (SPL), and soft SPL. An episode is considered successful when the agent issues the stop action within 1 meter of the goal. SPL [7] is defined as $\text{SPL}_\text{i} = S_i \cdot l_i / \max(p_i, l_i)$, where i denotes the index of the episode, $S = 1$ when the episode is successful and $S = 0$ otherwise, l denotes the length of the shortest path between the agent and the audio goal, and p denotes the length of the actual path taken by the agent in the episode. Soft SPL is a variation of SPL where $S_i = 1$ for all i.

TABLE I: Results of the AudioGoal navigation experiment. Our model strongly outperforms existing methods.

	SR ↑	SPL ↑	Soft SPL ↑
Random	0.01	0.07	0.12
DDPPO [21]	0.82	0.63	0.66
Direction Follower [9]	0.67	0.50	0.48
Beamforming [48]	0.02	0.01	0.24
Gan et al. [12]	0.63	0.53	0.68
AFP w/ predicting max	0.54	0.34	0.38
AFP w/o vision	0.84	0.71	0.72
AFP (Ours)	**0.91**	**0.76**	**0.75**

We compare with the following models: **DDPPO** [21]: an end-to-end reinforcement learning policy trained with distributed proximal policy optimization. **Direction Follower** [9]: this model predicts the direction of the audio goal and navigates with the same waypoint planner. We stop the agent automatically when it is within a 1-meter radius of the goal. **Gan et al.** [12]: this model predicts the (x,y) location of the audio source and navigates using a waypoint planner. The agent stops whenever it reaches its predicted location or the closest navigable point. **Beamforming** [48]: classical beamforming method that calculates the direction of arrival of the sound and navigates with the same waypoint planner.

To further justify our model design choice, we also compare with the following ablations of our own model. **AFP w/ predicting max**: this model does not predict the whole acoustic field. Instead, it predicts a single point that represents the highest point of the local acoustic field. **AFP w/ audio-only**: this model only takes in the audio input, which tests whether the full model uses the visual information when predicting the acoustic field.

Results are shown in Tab. I. Our model strongly outperforms all baselines and ablations. Compared to DDPPO, our model is more efficient due to its hierarchical nature since the DDPPO model often gets stuck with obstacles and corners. Direction Follower and Gan et al. predict the goal direction/location directly, which is however ill-posed when goals are in some other room at a distance from the robot. As a result, their navigation performance is also pretty poor. For the Beamforming baseline, similar to ours, it also predicts the local direction of arrival of the sound; however, since it is not robust to reverberation and noise, it performs poorly. Lastly, the two ablations perform comparably to baselines but also underperform the full model, showing it is both beneficial to predict the full acoustic field and leverage visual sensors to understand the environment.

We show the comparison of trajectories with these baselines in the same episode in Fig. 5, where our model is

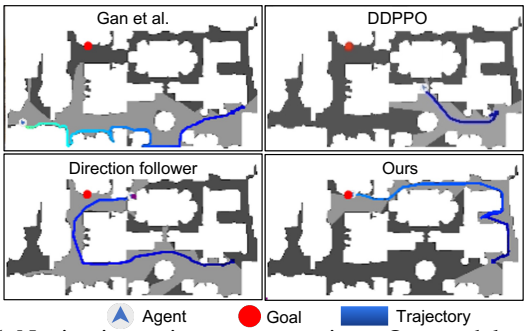

Fig. 5: Navigation trajectory comparison. Our model successfully navigates to the source while other baselines fail due to either getting stuck or navigating in the wrong direction.

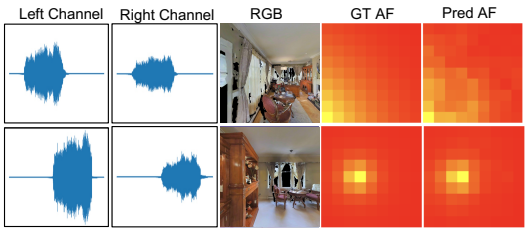

Fig. 6: Visualization of acoustic field prediction within the same episode. Top row: when the robot is still far from the goal. Bottom row: when the robot is right next to the goal. Our model predicts accurately in both cases.

more efficient in reaching the goal. We also visualize the acoustic fields of both the ground truth and prediction in Fig. 6. Initially, the model predicts high values at the corner (in the direction of the goal), and as the agent gets close to the goal, it predicts high values at the center of the field.

B. Experiment 2: Acoustic Field Prediction on Real Data

Here we evaluate our frequency-adaptive acoustic field prediction (FA-AFP) model on the collected real acoustic field data. We compare our method to the random baseline and ablations of our approach. We consider three ablations: "All-freq AFP" uses all frequencies for prediction. "Best-freq AFP" uses the best frequency band shown in Fig. 4 and "Highest-energy AFP" uses the band where the received audio has the highest energy. We measure the performance of different prediction errors with the angle and distance of the predicted max location on the acoustic field. We train our models on 73 sounds and test on 7 unheard sounds.

The results are shown in Tab. II. We show that compared to the random prediction, our All-freq AFP model reduces the prediction error drastically. If we always use the best frequency for prediction, it helps lower the angle error a bit but not the distance error. Using the frequency band with the highest energy brings down the prediction error more. Our frequency-adaptive prediction model (FA-AFP) improves the performance even further, showing the importance of intelligently selecting a frequency band for prediction.

In Fig. 7, we show examples of the collected acoustic field and the predicted acoustic field for multiple directions and sounds. Note that the acoustic field is only sampled at a 3×3 grid centered at the robot to reduce the cost of collection. Our predictions are consistently accurate across examples.

TABLE II: Results for testing on real acoustic field data.

	Angle ↓	Distance ↓
Random	1.57	1.45
All-freq AFP	0.22	0.74
Best-freq AFP	0.20	0.74
Highest-energy AFP	0.04	0.70
FA-AFP (Ours)	**0.04**	**0.63**

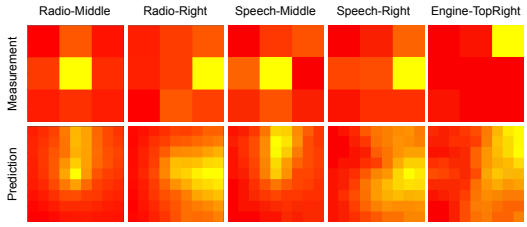

Fig. 7: Acoustic field predictions on real data. The real data is measured with a lower resolution. We show the prediction and measurement for multiple sounds and directions. Our model predicts all of these cases accurately.

C. Experiment 3: Real Robot Navigation

Finally, to validate the whole navigation pipeline, we deploy our navigation policy on the real robot platform (described in Sec. V). When deploying on the real robot, one thing that differs from the previously collected real data is that the robot also makes some low-frequency noise while running. To address this issue, we collect recordings of the robot noise and perform data augmentation by adding the noise to the received sound during training to improve the model performance.

We conduct 20 navigation examples with various source/receiver distances and directions and show that our robot can navigate the sounding object with a 75% success rate. See the supplementary video for both success and failure cases. We also tried to deploy the best-performing baseline DDPPO, which however failed all the test scenarios, which is likely due to the significant physical sim2real gap since that model trains with RL end-to-end. We show one navigation step example in Fig. 1, where the model predicts the acoustic field correctly.

VII. Conclusion

We systematically evaluate the sim2real acoustic gap with a proposed acoustic field prediction task. We further design a frequency-adaptive strategy to mitigate sim2real errors. We validate our model on both the Continuous AudioGoal navigation benchmark and collected real measurements. Lastly, we build a robot platform and show that we can successfully transfer the policy to the real robot.

While this work represents an exciting first step, it also has some limitations. First, the validation and test data were collected within the same environment, leaving the generalization to novel acoustic environments yet to be explored. Second, we assume the sound sources to be static, which may not hold in all cases, calling for new solutions to address dynamic objects.

Acknowledgements: UT Austin is supported in part by NSF CCRI and IFML NSF AI Institute. KG is paid as a research scientist by Meta.

REFERENCES

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Towards the robust-perception age," *IEEE Transactions on Robotics 32 (6) pp 1309-1332*, 2016.

[2] M. Deitke, D. Batra, Y. Bisk, T. Campari, A. X. Chang, D. S. Chaplot, C. Chen, C. P. D'Arpino, K. Ehsani, A. Farhadi, L. Fei-Fei, A. Francis, C. Gan, K. Grauman, D. Hall, W. Han, U. Jain, A. Kembhavi, J. Krantz, S. Lee, C. Li, S. Majumder, O. Maksymets, R. Martín-Martín, R. Mottaghi, S. Raychaudhuri, M. Roberts, S. Savarese, M. Savva, M. Shridhar, N. Sünderhauf, A. Szot, B. Talbot, J. B. Tenenbaum, J. Thomason, A. Toshev, J. Truong, L. Weihs, and J. Wu, "Retrospectives on the embodied AI workshop," 2022.

[3] 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.

[4] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *CVPR*, 2017.

[5] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A platform for embodied AI research," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9338–9346.

[6] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 4247–4258.

[7] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "Objectnav revisited: On evaluation of embodied agents navigating to objects," *arXiv:2006.13171*, 2020.

[8] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "Poni: Potential functions for objectgoal navigation with interaction-free learning," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 868–18 878.

[9] C. Chen, S. Majumder, Z. Al-Halah, R. Gao, S. K. Ramakrishnan, and K. Grauman, "Learning to set waypoints for audio-visual navigation," in *ICLR*, 2021.

[10] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural SLAM," in *ICLR*, 2020.

[11] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *ECCV*, 2020.

[12] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *ICRA*, 2020.

[13] P. Anderson, A. Shrivastava, J. Truong, A. Majumdar, D. Parikh, D. Batra, and S. Lee, "Sim-to-real transfer for vision-and-language navigation," in *CoRL*, 2020.

[14] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3803–3810.

[15] A. Kadian, J. Truong, A. Gokaslan, A. Clegg, E. Wijmans, S. Lee, M. Savva, S. Chernova, and D. Batra, "Sim2real predictivity: Does evaluation in simulation predict real-world performance?" in *RA-L*, 2020.

[16] R. Gao, H. Li, G. Dharan, Z. Wang, C. Li, F. Xia, S. Savarese, L. Fei-Fei, and J. Wu, "Sonicverse: A multisensory simulation platform for embodied household agents that see and hear," in *ICRA*, 2023.

[17] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in *CVPR*, 2021.

[18] C. Chen, C. Schissler, S. Garg, P. Kobernik, A. Clegg, P. Calamia, D. Batra, P. W. Robinson, and K. Grauman, "Soundspaces 2.0: A simulation platform for visual-acoustic learning," in *NeurIPS*, 2023.

[19] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015.

[20] J. Fuentes-Pacheco, J. R. Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, pp. 55–81, 2012.

[21] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," in *ICLR*, 2020.

[22] C. Gan, J. Schwartz, S. Alter, D. Mrowca, M. Schrimpf, J. Traer, J. De Freitas, J. Kubilius, A. Bhandwaldar, N. Haber, M. Sano, K. Kim, E. Wang, M. Lingelbach, A. Curtis, K. Feigelis, D. M. Bear, D. Gutfreund, D. Cox, A. Torralba, J. J. DiCarlo, J. B. Tenenbaum, J. H. McDermott, and D. L. K. Yamins, "ThreeDWorld: A platform for interactive multi-modal physical simulation," in *NeurIPS Datasets and Benchmarks Track*, 2021.

[23] K. Nakadai and K. Nakamura, "Sound source localization and separation," *Wiley Encyclopedia of Electrical and Electronics Engineering*, 1999.

[24] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, 2017.

[25] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *AAAI*, 2000.

[26] X. Wu, H. Gong, P. Chen, Z. Zhong, and Y. Xu, "Surveillance robot utilizing video and audio information," *Journal of Intelligent and Robotic Systems*, 2009.

[27] J. Qin, J. Cheng, X. Wu, and Y. Xu, "A learning based approach to audio surveillance in household environment," *International Journal of Information Acquisition*, 2006.

[28] T. Yoshida, K. Nakadai, and H. G. Okuno, "Automatic speech recognition improved by two-layered audio-visual integration for robot audition," in *2009 9th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2009, pp. 604–609.

[29] X. Alameda-Pineda and R. Horaud, "Vision-guided robot hearing," *The International Journal of Robotics Research*, 2015.

[30] R. Viciana-Abad, R. Marfil, J. Perez-Lorenzo, J. Bandera, A. Romero-Garces, and P. Reche-Lopez, "Audio-visual perception system for a humanoid robotic head," *Sensors*, 2014.

[31] J. M. Romano, J. P. Brindza, and K. J. Kuchenbecker, "Ros open-source audio recognizer: Roar environmental sound detection tools for robot programming," *Autonomous robots*, 2013.

[32] T. Yoshioka and T. Nakatani, "Dereverberation for reverberation-robust microphone arrayss," in *EUSIPCO*, 2013.

[33] P. P. Rao and A. R. Chowdhury, "Learning to listen and move: An implementation of audio-aware mobile robot navigation in complex indoor environment," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 3699–3705.

[34] T. Jenrungrot, V. Jayaram, S. Seitz, and I. Kemelmacher-Shlizerman, "The cone of silence: Speech separation by localization," in *Advances in Neural Information Processing Systems*, 2020.

[35] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson env: Real-world perception for embodied agents," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9068–9079.

[36] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," in *2017 International Conference on 3D Vision (3DV)*, 2017, pp. 667–676.

[37] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "AI2-THOR: An Interactive 3D Environment for Visual AI," *arXiv*, 2017.

[38] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IROS*, 2017.

[39] J. Vacaro, G. Marques, B. Oliveira, G. Paz, T. Paula, W. Staehler, and D. Murphy, "Sim-to-real in reinforcement learning for everyone," *LARS-SBR-WRE*, 2019.

[40] K. Kristinsson and G. A. Dumont, "System identification and control using genetic algorithms," *IEEE Transactions on Systems, Man, and Cybernetics, 22(5)*, 1992.

[41] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, 2020.

[42] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. De Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, "The replica dataset: A digital replica of indoor spaces," *arXiv*, 2019.

[43] A. Younes, D. Honerkamp, T. Welschehold, and A. Valada, "Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds," *RAL*, 2021.

[44] C. Cao, Z. Ren, C. Schissler, D. Manocha, and K. Zhou, "Interactive

sound propagation with bidirectional path tracing," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–11, 2016.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[46] J. A. Sethian, "Fast marching methods," *SIAM review*, vol. 41, no. 2, pp. 199–235, 1999.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
[48] S. Lapp, T. Rhinehart, L. Freeland-Haynes, J. Khilnani, A. Syunkova, and J. Kitzes, "Opensoundscape: An open-source bioacoustics analysis package for python," *Methods in Ecology and Evolution 2023*, 2023.