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

Changan Chen[1*], Jordi Ramos[1*], Anshul Tomar[1*], Kristen Grauman[1,2]

1. Extended Abstract

(For full paper and supplemantry video, see vision.cs.utexas.edu/projects/sim2real)

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) [3] with Lidar sensors. This approach however is limited to geometry planning and does not reason well about the semantics of the scene. In addition, building robots and reproducing experiments are expensive for real robots.

To address these issues, recent years have witnessed research approaching the navigation problem from a vision-centric perspective, i.e., the robot mainly uses visual sensors to perceive the scene [10]. This approach has demonstrated a lot of success in photorealistic real-scanned environments, which allow fast experimentation and replicable research. Various forms of tasks have been proposed in this domain including using egocentric vision to travel to a designated point in an unfamiliar environment [13, 19, 22], search for a specified object [2, 5], or explore and map a new space [4, 8, 17]. Some other works further explore expanding the sensory suite of the navigating agent to include hearing as well. In particular, the AudioGoal task [6, 11] 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 photorealistic simulation environments, some work explores transferring the learned policy to the real world by bridging the gap between the simulation and the real world [1, 14, 16, 21]. Recent work [12] does sim2real transfer for audio-visual navigation with data augmentation

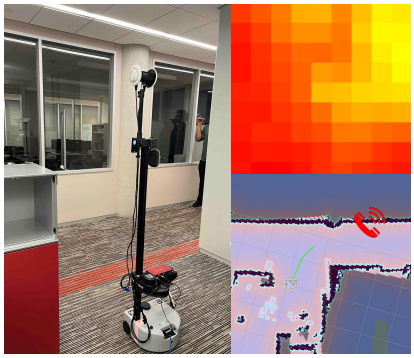

Figure 1. Our robot predicts an acoustic field with a frequency-adaptive model and navigates to locate the sound source.

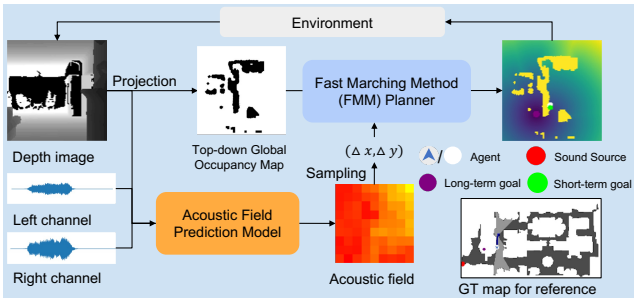

Figure 2. Navigation pipeline. The model first predicts the acoustic field, samples the peak as the long-term goal, and navigates toward the goal with a path planner.

however without further investigating the acoustic gap. The sound differs from light in that it spans a wide range of frequencies, which is one of the main barriers to sim2real transfer. In this work, we perform a systematic evaluation of the acoustic gap and propose a solution to bridge that gap.

State-of-the-art approaches in audio-visual navigation rely on reinforcement learning to train the navigation policy end-to-end [6, 7], which is not only hard to interpret but also impractical to generalize to the real world directly due to various sim2real gaps. Recent visual navigation work has shown success in sim2real transfer with hierarchical models [1, 17], which typically consist of a high-level path planner and low-level motion planner. This design helps abstract away some of the low-level physical discrepancies.

Inspired by such methods, we design a modular approach

* indicates equal contribution, sorted in alphabetical order

[1]University of Texas at Austin, [2]FAIR, Meta AI

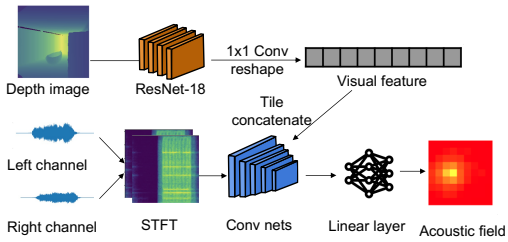

Figure 3. Acoustic field prediction model. The model first extracts audio and visual features, and then tiles, and concatenates both features to predict the acoustic field.

	SR ↑	SPL ↑	Soft SPL ↑
Random	0.01	0.07	0.12
DDPPO [20]	0.82	0.63	0.66
Direction Follower [8]	0.67	0.50	0.48
Beamforming [15]	0.02	0.01	0.24
Gan et al. [11]	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**

Table 1. Results of the AudioGoal navigation experiment. Our model strongly outperforms the SOTA method on this benchmark.

to ease the transfer from the 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, they are free to compute in simulation. We first build an audio-visual model (see Fig. 3) as the acoustic field predictor (AFP) and curate a large-scale acoustic field dataset on SoundSpaces 2.0 [9], the state-of-the-art audio-visual simulation platform. We show that this approach outperforms existing methods on the Continuous AudioGoal navigation benchmark (see Tab. 1).

After validating the proposed approach in simulation, we 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 [18]. Given this observation, we focus on evaluating how the sim2real error changes as a function of 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

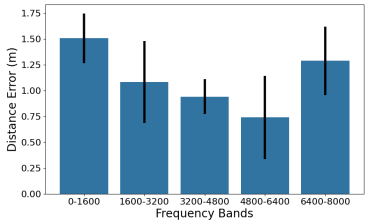

Figure 4. Sim2real error as a function of frequencies. We report the mean and standard deviation of distance errors between the predicted and the ground truth peak locations.

	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**

Table 2. Results for testing on real acoustic field data.

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 (see Fig. 4). However, simply taking the best frequency band doesn't work for all sounds since different sounds have different spectral distributions. To address this issue and make the model aware of the spectral difference, we propose a novel frequency-adaptive prediction strategy, that intelligently selects the best frequency sub-band based on measured errors as well as the received spectral distribution to predict the acoustic field. 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 (see Tab. 2). 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. We perform a systematic evaluation of the sim2real 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 principal solution to the sim2real transfer problem for audio-visual navigation.

References

[1] Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. Sim-to-real transfer for vision-and-language navigation. In *CoRL*, 2020. 1

[2] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects, 2020. 1

[3] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Towards the robust-perception age. *arXiv*, 2016. 1

[4] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *ICLR*, 2020. 1

[5] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *Advances in Neural Information Processing Systems*, pages 4247–4258, 2020. 1

[6] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020. 1

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

[8] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *ICLR*, 2021. 1, 2

[9] 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. In *NeurIPS*, 2023. 2

[10] Matt Deitke, Dhruv Batra, Yonatan Bisk, Tommaso Campari, Angel X. Chang, Devendra Singh Chaplot, Changan Chen, Claudia Pérez D'Arpino, Kiana Ehsani, Ali Farhadi, Li Fei-Fei, Anthony Francis, Chuang Gan, Kristen Grauman, David Hall, Winson Han, Unnat Jain, Aniruddha Kembhavi, Jacob Krantz, Stefan Lee, Chengshu Li, Sagnik Majumder, Oleksandr Maksymets, Roberto Martín-Martín, Roozbeh Mottaghi, Sonia Raychaudhuri, Mike Roberts, Silvio Savarese, Manolis Savva, Mohit Shridhar, Niko Sünderhauf, Andrew Szot, Ben Talbot, Joshua B. Tenenbaum, Jesse Thomason, Alexander Toshev, Joanne Truong, Luca Weihs, and Jiajun Wu. Retrospectives on the embodied ai workshop, 2022. 1

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

[12] Ruohan Gao, Hao Li, Gokul Dharan, Zhuzhu Wang, Chengshu Li, Fei Xia, Silvio Savarese, Li Fei-Fei, and Jiajun Wu. Sonicverse: A multisensory simulation platform for embodied household agents that see and hear. In *ICRA*, 2023. 1

[13] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *CVPR*, 2017. 1

[14] Abhishek Kadian, Joanne Truong, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. Sim2real predictivity: Does evaluation in simulation predict real-world performance? In *RA-L*, 2020. 1

[15] Sam Lapp, Tessa Rhinehart, Louis Freeland-Haynes, Jatin Khilnani, Alexandra Syunkova, and Justin Kitzes. Opensoundscape: An open-source bioacoustics analysis package for python. *Methods in Ecology and Evolution 2023*, 2023. 2

[16] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3803–3810, 2018. 1

[17] Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18868–18878, 2022. 1

[18] Lauri Savioja and U Peter Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 2015. 2

[19] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9338–9346, 2019. 1

[20] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *ICLR*, 2020. 2

[21] Wenshuai Zhao, Jorge Pena Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: A survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 737–744, 2020. 1

[22] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J. Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*, 2017. 1