2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

The Un-Kidnappable Robot: Acoustic Localization of Sneaking People

Mengyu Yang[1*], Patrick Grady[1], Samarth Brahmbhatt[2],
Arun Balajee Vasudevan[3], Charles C. Kemp[4], James Hays[1]

***Abstract*— How easy is it to sneak up on a robot? We examine whether we can detect people using only the incidental sounds they produce as they move, even when they try to be quiet. To do so, we first collect a robotic dataset of high-quality 4-channel audio paired with 360° RGB data of people moving in different indoor settings. Using this dataset, we train models to predict if there is a moving person nearby and then their location using only audio. We implement our method on a robot, allowing it to track a single person moving quietly using only passive audio sensing. For demonstration videos, see our project page.**

I. INTRODUCTION

Advances in mobile robots have led to such platforms becoming increasingly common in everyday settings. With this popularity comes a rise in the coexistence of robots with people. Nowadays, it is not uncommon to see a last-mile delivery robot roaming a city sidewalk, an industrial robot navigating a warehouse floor, or a cleaning robot vacuuming in a home. As demand for robotic applications grows, being able to recognize people in the robots' proximity is a vital task to ensure safety. Object recognition in general has been well examined for image data [1], [2], [3], where humans are one of the object categories. Previous works have also specifically investigated person detection [4], [5], [6], [7], [8] by detecting the presence of people as well as localizing them. The large amount of existing research in this topic highlights the universality of person detection across different robotic applications.

Most person detection methods use vision- and spatial-based sensors. These include RGB [11], depth [12], 2D lasers [13], [14], 3D LIDAR [7], or combinations for multi-modal methods [15], [16]. While multi-modal models are beneficial for domain adaptation and improved performance over uni-modals, these methods learn a joint representation across all sensors that do not allow flexibility for variable sensor inputs. These models are therefore not robust against failures in real-world applications where sensors can fail. Aside from these common sensors, fewer works have examined the use of audio for person detection. Existing works only use audio to estimate the direction of arrival (DOA) [17], [18], [19], [20] which does not satisfy our definition of person detection. Others that do perform full localization tend to rely on easily detectable sounds like talking or audio from a loudspeaker [21], [22], [23], [24].

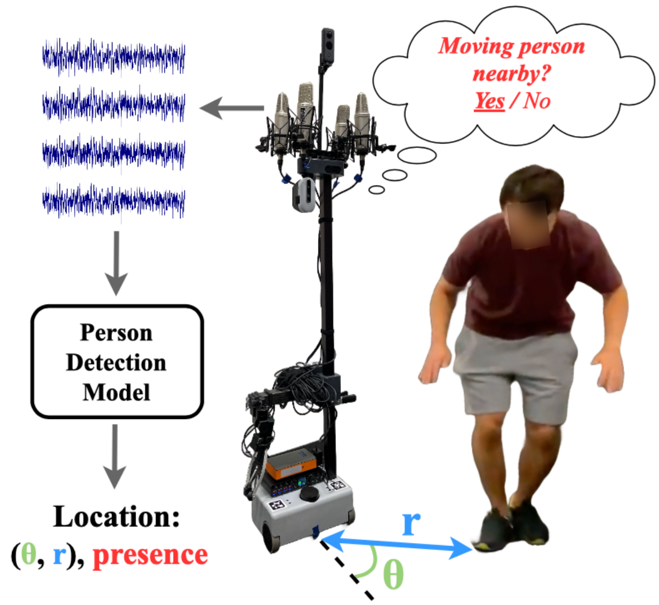


Fig. 1: Can we detect where people are based only on the subtle sounds they incidentally produce when they move, even when they try to be quiet? We collect a dataset of high-quality audio paired with 360° RGB data with different participants in multiple indoor scenes. We train models to localize a moving person based on audio only and implement it on a robot.

We argue that the subtle acoustics *incidentally* produced by people as they move around are an under-leveraged source of information that can be used for person detection. The incidental sounds we focus on are by nature noisy and weak. We demonstrate this by showing how other sound localization methods [25], [26] often fail to detect the sound source for this type of audio. Unlike other sound localization methods, ours relies on *passive observation* only. We only record the sounds already coming from the environment and do not produce any additional sounds to aid in detection. This is in contrast to other methods using active sensing, such as with ultrasonic sensors [27], echolocation [28], or the impulse response of a room [29].

Having an audio-only based method for person detection is an important step in the development of multi-modal person detection systems that are robust to failures. Should the sensors that many frameworks rely on fail or become unavailable (low-lit environments, occlusion handling, etc.), our method allows robots to fall back solely onto audio, a

*Corresponding author: my.yang@gatech.edu
[1]Mengyu Yang, Patrick Grady, and James Hays are with the Georgia Institute of Technology (Georgia Tech)
[2]Samarth Brahmbhatt is with Intel Labs
[3]Arun Balajee Vasudevan is with Carnegie Mellon University
[4]Charles C. Kemp worked for Georgia Tech and now works full-time for Hello Robot Inc., which sells the Stretch RE-1

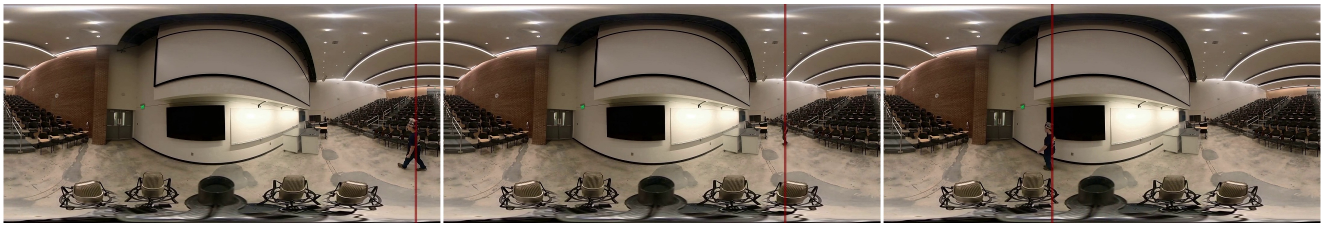


Fig. 2: Frames from the Robot Kidnapper dataset (static robot). The participant wears a hat with ArUco markers [9] used to calculate ground truth radial distance. The RGB frames are used to calculate the ground truth centroid of the person using DeepLabv3+ [10]. Only the audio is used during training. The vertical red lines are the angles predicted by our model in an unseen room. The participant is walking normally in these frames.

readily obtainable signal which is usually already onboard most hardware setups. And when interacting with robots, people should not be expected to intentionally create extra sounds to ensure nearby robots are aware of their location.

To evaluate our claims, we first collected a real-world dataset of different people moving around a robot in various indoor settings. Onboard the robot, we record 4-channel, high-quality audio along with paired 360° RGB data, which we process to obtain pseudo-labels for the person's location relative to the robot. We name this the Robot Kidnapper dataset (Fig. 2, 3a) and provide more details in Section III.

We then use this dataset to learn person detection based on the incidental and often subtle sounds created by people as they move around. We show that our models are able to localize people both when the robot is stationary and when the robot is moving, a more difficult task due to the additional self-noise of the robot. We also implement our model on a real robot to demonstrate robotic human awareness using only audio. Overall, we present the following contributions:

1) A public dataset of synchronized, high-quality, 4-channel audio and 360° RGB data of different participants in multiple indoor scenes. **2**) Experimental evaluations of person detection using the subtle, incidental sounds of people moving around. **3**) Allowing real robots to track people using only the sounds of them moving.

II. RELATED WORK

A. Human Detection with Visual Perception

Since perceiving humans enables a large number of downstream tasks, 'person' or 'human' is included as a category in most image-based object detection [30], [31] and segmentation [32] models. Autonomous cars use LIDAR sensors optionally combined with RGB cameras to detect pedestrians [33], [34] and forecast their future behaviour [35]. Additionally, surveillance systems use RGB or infrared images to detect [36], [37], [38], identify [39], and localize humans [40] and objects [41]. In settings where it can be assumed that any disturbances are caused by humans, human detection is performed through anomaly detection algorithms. For example, laser/ultraviolet beam breakers or proximity sensors are used on factory floor automated assembly lines [42], while some intrusion detection security systems [43] use audio. In contrast to these works, our paper focuses solely on passively sensing audio signals to not only detect but also localize humans.

B. Audio-Based Perception for Robots

Audio has been used for robotic tasks involving both non-human and human interactions such as self-localization [44], [45], robotic pouring [46], and navigation using ambient sounds [47]. Sound source localization systems like [48], [49], [50] usually assume that the source emits loud, obvious sounds (*e.g.* beeps, music, speech). But as Fig. 5 shows, the more subtle sounds we focus on have different acoustic characteristics which these algorithms may not successfully generalize to. [17], [18], [19], [20] all examine human detection but only estimates the direction of arrival. In contrast, we perform full 2D localization by also predicting the radial distance of the person from the robot. And while Sasaki et al. [50] performs 3D localization, they assume a static sound source with a moving robot. In our work, we assume both the robot and sound source can be moving at the same time.

III. DATASET

To train our models to detect people based on the incidental sounds that they produce, we collected the Robot Kidnapper dataset. This dataset contains high-quality 4-channel audio recordings paired with 360° RGB video from the robot's egocentric point of view (Fig. 2). The person's position was annotated in coordinates relative to the robot. We collected data in 8 rooms across 4 buildings. To account for the potential impacts that physical properties of a room may have, the selected rooms vary in terms of size (small study room, large lecture hall, etc.) and material (concrete floors, carpeted floors, glass walls, etc.).

A. Human Presence Recordings

The Robot Kidnapper dataset captures 12 participants in a range of environments performing a variety of actions. This *stress-tests* the performance of our algorithm across diverse behaviors. Participants were prompted to perform 4 different actions during data capture to capture a wide range of sounds:

- *Stand still:* Participants were asked to stand in place for 5 seconds before taking 1-2 steps to a different spot and repeating the procedure.
- *Walk quietly:* Participants were prompted to move during the entire recording but to focus on minimizing any sounds that they produced.
- *Walk normally:* Participants were prompted to walk at their normal speed and volume

- *Walk loudly:* Here, participants were prompted to walk more loudly, which they accomplished by dragging their feet or stomping.

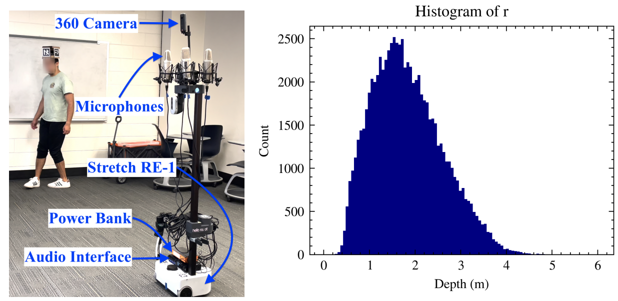
During data collection, the hardware was mounted to a Stretch RE-1 mobile manipulator robot. However, the robot produces sounds during movement, such as humming and clicking from wheel motors or rustling as the robot traverses bumps on the ground. We examine whether our methods can learn under the more difficult setting of detecting a moving person with these additional noises. For all 4 actions, we collected data under a *static robot* and *dynamic robot* condition. During the *static* recordings, the robot was turned on but remained stationary. In the *dynamic* recordings, the robot was teleoperated from outside the room and driven (translation and rotation) along a random path, with the aim of uniformly covering the entirety of the room. The robot moved at a translational and rotational velocity of approximately 0.25 m/s and 0.17 rad/s, respectively. Participants were alone in the room with the robot. As our intention is to examine how well our models can detect people from only incidental sounds, non-incidental sounds such as talking were cropped out during post-processing. All recordings contain a single participant only. The data collection was approved by an Institutional Review Board (IRB) and participants gave informed consent and were compensated for their time. All combined, the human presence recordings total to approximately 8 hours, evenly split between all actions, robot conditions, and rooms.

B. Empty Room Recordings

In addition to recording the sounds of human presence, we also recorded audio of the 8 rooms used in the dataset when it was empty. This empty room data helps our model be able to distinguish whether or not there is a moving person in the robot's vicinity. The empty room recordings are collected on the same robot setup as the human recordings. It is also split between *static* and *dynamic*. This empty room dataset is approximately 5 hours in length. We then collected a secondary empty room dataset without the Stretch RE-1 consisting only of short recordings. The *Empty Augmentation* dataset was collected in 26 rooms across 6 buildings on Georgia Tech's campus. In each room, audio of the empty room was recorded from 2 different positions for 2 minutes each, resulting in around 1.5 total hours of audio. The audio from this dataset is used for data augmentation, which is described in Section IV-B.

C. Person Location Labels

Training a person detection model requires ground truth labels for the location of the person relative to the robot. In our dataset, we annotate the person's position on the ground plane in polar coordinates. Specifically, we annotate the azimuthal angle θ of the person relative to the robot's forward vector, and radial distance r of the person relative to the robot's origin. To label θ, we developed an approach based on a semantic segmentation model. Specifically, we use a pre-trained DeepLabv3+ [10] model to generate a



(a) (b)

Fig. 3: **(a)** Dataset capture setup. **(b)** Distribution of radial distances between the robot and person in the dataset.

mask of the person from an RGB frame and calculate the centroid of the person (x, y). We then encode x using cyclical features, where W is the width of the frame.

$$\theta_{sin} = sin(2\pi x/W) \quad (1)$$

$$\theta_{cos} = cos(2\pi x/W) \quad (2)$$

To label r, participants wore a hat with ArUco markers [9]. Patches of the 360° RGB frame were re-projected using a pinhole camera model and an ArUco pose estimator was run on these frames. Due to factors such as motion blur, lighting, and low resolution at further distances, ArUco markers were detected in only 74% of all the frames in the dataset. The distribution of r is shown in Fig. 3b. Each dataset sample consists of a 1s clip of audio and corresponding video. We use the first frame of the video clip to extract location labels. We sample overlapping clips at 4Hz. We note that while robust RGB-D cameras are readily available, they are limited to a narrow field of view and not suitable for our 360° data.

D. Hardware

All audio recordings were done at 44.1 kHz using 4 RØDE NT2-A microphones connected to a MOTU M6 audio interface. The polar pattern for all 4 microphones was set to cardioid with the front side facing outwards. To ensure that we capture the maximum amount of information from the audio, the microphone gain was adjusted to the highest setting, the PAD was set to 0 dB and the high-pass filter was set to flat. An Insta360 ONE X2 was used for the 360° video recordings, with the video frames processed into 1440x720 resolution using equirectangular projection. The hardware was mounted on the Stretch RE-1 robot (Fig. 3a). For the Empty Augmentation dataset, a standard tripod was used.

IV. METHODOLOGY

We examine acoustic localization of people using only the incidental sounds produced by their moving presence. We design and train models on our dataset of high-quality multi-channel audio paired with 360° RGB data from which we extract location labels. While we collected data of people standing still as well, our paper only analyzes the moving actions (quiet, normal, loud). We train our models using leave-one-out cross validation across all 8 rooms in the dataset. All results are from averaging test performance across the 8 unseen rooms.

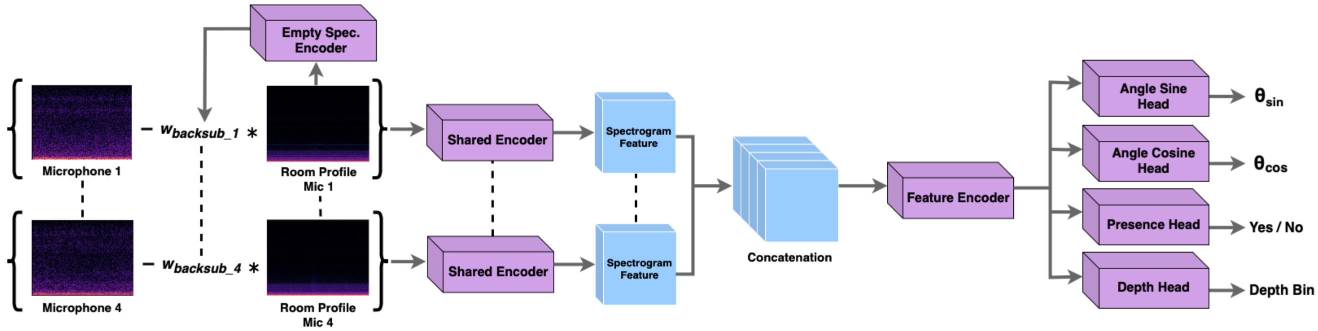


Fig. 4: Diagram of our model architecture. We perform background subtraction (Sec. IV-A) on input spectrograms before passing them through a spectrogram encoder with shared weights. The resulting features are concatenated and passed through the feature encoder based on the ASPP module [10]. The output is fed to 4 linear layer heads for the prediction tasks.

A. Background Subtraction

Given the weak-signal nature of the data, our models must be robust to the ambient noises in different rooms. As seen in Fig. 5, aside from loud walking, the other actions are difficult to distinguish from that of an empty room. This shows that most of the audio signals consist of background noise that our model must learn to ignore. To do so, we use a simple form of spectral subtraction. Right before recording in each room, we first collected 20s of empty room audio with either a static or dynamic robot. We split that audio into non-overlapping 1s clips, compute their spectrograms, and then take the average spectrogram. This average spectrogram, S_{empty}, is the empty room profile. Now given an input spectrogram of audio from the same room, S_{in}, we perform *background subtraction* by computing $S_{final} = S_{in} - w_{backsub} * S_{empty}$. $w_{backsub}$ is a scalar weight that the empty spectrogram encoder learns (Fig. 4). We clamp the values to [0, 1].

B. Empty Room Augmentation

Training our models to adapt to different background noises requires a wide variety of rooms to be seen during training. To supplement the 8 rooms we collected data in, we synthetically create additional rooms during training. Since audio is additive, we can inject additional noise into an audio recording and place the sounds of a moving person in a new room, one which contains a combination of ambient noises from two different rooms. This is the purpose of the Empty Augmentation dataset from Section III-B. Given an audio clip from the Robot Kidnapper dataset, $x_r(t)$, and an audio clip of equal length of an empty room from the Empty Augmentation dataset, $x_{aug}(t)$, we first normalize both clips to an RMS of 0.02. Then, we calculate a linear combination of the two waveforms: $x_{syn}(t) = (1-w_{aug}) * x_r(t) + w_{aug} * x_{aug}(t)$. w_{aug} is a scalar which we tune as a hyperparameter. We normalize $x_{syn}(t)$ again before processing its spectrogram.

For background subtraction, the same procedure described in Section IV-A is used to calculate empty room profiles for both the natural and synthetic empty room, resulting in S^{nat}_{empty} and S^{syn}_{empty}. To compute the final empty room profile, the same weighting factor w_{aug} is used: $S_{empty} = (1 - w_{aug}) * S^{nat}_{empty} + w_{aug} * S^{syn}_{empty}$. S_{empty} is then used for background subtraction.

C. Models

Architecture: We adapt the audio encoder architecture from Vasudevan et al. [51] for our person detection task. The network takes a 1s clip of audio in the form of a spectrogram. To generate the spectrogram, we first normalize the raw waveform to a constant RMS value of 0.02. The waveform is then fed through a Short-Time Fourier Transform (STFT) with a window size of 512 and a hop length of 128 and then converted to the log scale. This results in a $[2, 257, 345]$ spectrogram for each microphone, where the 2 channels correspond to the real and complex components.

As seen in the architecture diagram in Fig. 4, we first subtract the empty room profile spectrogram from the input spectrogram (Section IV-A) for each microphone before passing them individually through a spectrogram encoder with shared weights, consisting of 4 strided convolutional layers. The output is a $[256, 60, 120]$ feature map for each spectrogram. These features are then concatenated in the channel dimension to form a $[256n, 60, 120]$ feature map, where n is the number of microphones being used, before being passed through the feature encoder. The feature encoder is an Atrous Spatial Pyramid Pooling (ASPP) module [10], which [51] found to be a powerful audio encoder for spatial audio tasks. We refer readers to [10], [51] for a more detailed description of the architecture. The output of the feature encoder is a $[1, 240, 480]$ feature map. The flattened feature map is passed through 4 task-specific decoders, each being a linear layer. Each decoder predicts one of the following:

Azimuthal angle prediction: $\hat{\theta}_{sin}$ and $\hat{\theta}_{cos}$ are each predicted by a decoder. We clamp the predictions to $[-1, 1]$ and then decode back to the pixel coordinate $\hat{x} = tan^{-1}(\hat{\theta}_{sin}/\hat{\theta}_{cos})$. We then apply a L1 loss between $\hat{x}$ and x. The loss values for empty room training samples are ignored so the model does not try to learn to predict the location of a non-existent person.

Radial distance prediction: We frame radial distance r estimation as a binary classification task by predicting if a person is within 1.7m of the robot, which is the median of the distribution in Fig. 3b. We train using a binary cross-entropy loss and losses for empty room samples are ignored.

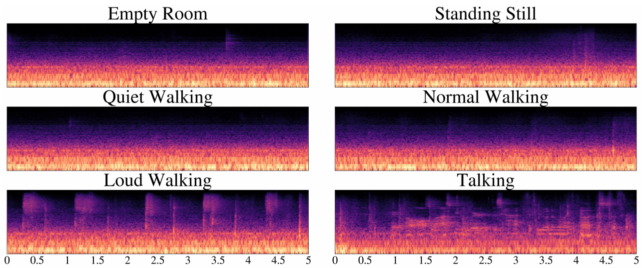


Fig. 5: Log spectrograms for all categories along with regular talking. No talking is used in our work, but we show the spectrogram as reference for a common sound source used in localization. All recordings were taken in the same room during the same recording session.

Motion presence prediction: The model learns a binary classification task of whether or not a person is moving in the room. A binary-cross entropy loss is used.

Training: We use the Adam [52] optimizer with a learning rate of 10^{-4}, momentum of 0.9, and weight decay of 10^{-3}. We train the entire model using a multi-task framework on a single NVIDIA A40 GPU. The model has 8.37M parameters.

V. Experiments

We present the person detection performance of our model trained on the Robot Kidnapper dataset. Performance is broken down into 3 tasks: θ, r, and moving presence prediction. All results are the average performance across the 8 test room folds. We then compare against other methods and perform ablation studies to validate our model design. Finally, we demonstrate our model in the real world by implementing it on the Stretch RE-1 robot.

A. Model Comparisons

For the angle prediction, we compare our model with GCC-PHAT [25], a commonly used handcrafted feature, and StereoCRW [26], an unsupervised method that learns spectrogram representations. Both methods estimate the time delay between two stereo channels from which the DOA can be calculated. For StereoCRW, we run inference with the provided pre-trained model weights which had been trained on significantly more data and demonstrates generalization capabilities. Both comparison models are designed for 2 microphones which can only predict the direction of sound within the range [-90°, 90°]. To compare with our 360° method, we use an oracle which always selects the pair of microphones (front 2 or back 2 microphones) facing the person. We also compare against a naive oracle method, constant front, which always predicts 0° (straight ahead) relative to the microphone pair selected by the oracle.

Both radial distance and moving presence prediction are binary classification tasks, which we compare against chance. While Chen et al. [47] examines a related task of estimating distance to nearby walls based on ambient sounds of the room, they focus on non-sound producing objects. Meanwhile, we treat ambient sounds as noise and instead focus on the subtle sounds that are present within.

TABLE I: Mean absolute error (MAE) in degrees for azimuthal angle prediction of our model and comparison methods across the 3 actions divided by static (Sta.) and dynamic (Dyn.) robot. Our Base 4 Mics model is trained without background subtraction (Sec. IV-A) and empty room augmentation (Sec. IV-B).

		Quiet		*Normal*		*Loud*	
Category	Model	Sta.	Dyn.	Sta.	Dyn.	Sta.	Dyn.
Random	Uniform 360°	90	90	90	90	90	90
Oracle Mic Pair	Constant Front	50	43	50	46	50	43
	GCC-PHAT [25]	44	47	45	43	46	47
	StereoCRW [26]	52	46	51	48	37	34
Ours	1 Mic	67	75	64	71	64	74
	2 Mics	37	54	37	48	36	47
	Base 4 Mics	47	55	50	48	49	47
	4 Mics	**21**	**26**	**22**	**24**	**19**	**22**

B. Azimuthal Angle Prediction

We evaluate 3 variations of our model, each trained on a different number of microphones: all 4 microphones, the front 2 microphones, and a single front microphone. We calculate the mean absolute error (MAE) in degrees to measure angle prediction performance. We show separate metrics for inference on static and dynamic robot recordings.

Looking at Table I, we notice that both GCC-PHAT and StereoCRW have similar performance as the naive constant front method. This supports our claim that the subtle sounds we focus on are difficult for previous sound localization methods. An exception is StereoCRW on the loud category, which performs noticeably better than the other methods being compared. Fig. 5 qualitatively shows how the acoustic characteristics of the loud category is similar to that of talking. Since other sound localization methods tend to focus on relatively prominent sources of sounds, it makes sense that a similar category like loud walking is detectable as well. But for the more subtle actions, other methods are unable to pick out the useful sounds from the background noise.

Moving on to our models' performance, our 4-microphone model significantly outperforms all other methods, with both GCC-PHAT and StereoCRW having approximately twice the MAE across all categories. This shows that the incidental sounds created by a moving person provides a rich source of cues for person detection. Performance on static robot recordings are generally better than dynamic robot, suggesting that the added self-noise of a moving robot complicates the already difficult task of detecting these subtle sounds.

Looking next at our 2-microphone model, it outperforms the other methods for all static recordings. Recall that our 2-microphone model is at a disadvantage compared to the non-random methods we compare to, since those models have access to all 4 microphones and always picks the ideal pair facing the person. They only need to predict angles within the range [-90°, 90°] while our model, with access to only a fixed pair of microphones, has to predict the entire 360° range. This again highlights our model's ability to detect and localize the subtle, incidental sounds produced by people when they move, even under constrained input settings.

While methods using time difference estimation can only unambiguously predict the direction of sound within 180°

Overall Accuracy: 67%

Quiet		*Normal*		*Loud*	
Sta.	Dyn.	Sta.	Dyn.	Sta.	Dyn.
67	61	71	66	71	66

(a)

Overall Accuracy: 87%

NEGATIVE		POSITIVE					
Empty		*Quiet*		*Normal*		*Loud*	
Sta.	Dyn.	Sta.	Dyn.	Sta.	Dyn.	Sta.	Dyn.
81	85	89	80	95	92	96	94

(b)

TABLE II: **(a)** Binary classification accuracy (%) of predicting if a moving person's radial distance is above or below 1.7m. Results are separated by action and static (Sta.) and dynamic (Dyn.) robot recordings. Chance is 50%. **(b)** Binary (negative vs positive) classification accuracy (%) of predicting if there is a moving person in the room. We also separate results by action, static (Sta.), and dynamic (Dyn.) robot recordings within each class. Chance is 50%.

due to symmetry, the configuration of our microphones allows us to predict a wider angle range. We postulate that the cardioid polar pattern on our microphones breaks this symmetry by being more sensitive to the sounds coming from the front of the microphone versus the back. The model is able to learn from this subtle difference and determine which side the sound is coming from, even if the time difference between the two microphones is the same. Finally, we also train our model on audio from only 1 fixed microphone. Understandably, this variation performs worse than the non-random methods. But the 1-microphone model still performs better than chance, suggesting that it is able to pick up some cues from the mono audio.

C. Radial Distance Prediction

We also estimate the radial distance r which, combined with θ, gives us the location of the person. We frame this as a binary classification problem by setting the threshold to the median (1.7m) of the distribution (Fig. 3b) and predicting if the person is above or below that threshold. The model is trained on 4 microphones and performs better than chance (50%), shown in Table IIa. Quiet walking is the most difficult action for this task since it has the least amount of signal. However, performance seems to saturate at normal and loud walking, with similar performance between both actions.

D. Moving Presence Prediction

We also want to detect the presence of a moving person in addition to localizing them. We evaluate our models on the binary classification task of differentiating between the audio of a person present and moving in the room (positive) and that of an empty room (negative) in both static and dynamic robot recordings. The model is trained on 4 microphones and performs significantly better than chance (Table IIb). As expected, the louder actions have better performance, since there are more obvious signals for the model to detect.

E. Ablation Studies

We validate our model design by training a base, 4-microphone model with neither background subtraction nor

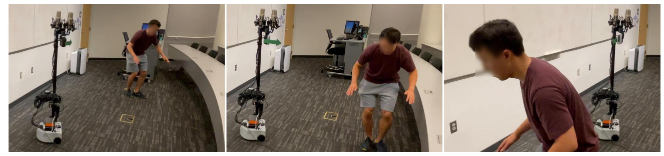


Fig. 6: We implement our trained model on the Stretch RE-1 robot to track a person using only the incidental sounds created as they move quietly. The robot pans the RealSense camera, with green arrow attached, to face where the model estimates the person to be (zoom in for best results).

empty room augmentation, shown in Table I. The base model has twice the MAE, demonstrating the necessity of these two features and the difficulty of the sounds we are learning from. Without additional methods to remove background noise, deep learning models have trouble detecting these subtle sounds. We do not decouple background subtraction and empty room augmentation because they complement each other, with the former allowing the model to adapt to different types of background noises while the latter provides more diverse training data for the model to actually learn the conditioning. Also, the 2- and 1-microphone models discussed in Section V-B can be seen as an ablation on the number of microphones.

F. Robotic Human Awareness

We implement our trained model on the Stretch RE-1 to demonstrate robotic human awareness. Using the same hardware setup and in an unseen room, we input the most recent 1s clip of audio into the model and use the predicted angle to pan the RealSense camera on the Stretch to face the person. During the pan, the model does not perform inference to avoid the sound of the motor interfering with angle estimation. On a RTX 2080, the model runs at 142Hz. Fig. 6 provides a demonstration of the robot pointing at the person. We do not use the RealSense data in any way to estimate the person's direction. Given the narrow field of view of the RealSense (58°), we evaluate our algorithm's performance by determining the success rate of the person being present in the frame right after panning. We asked a participant to move at each of the 3 speeds for 4 minutes and obtained the following success rate: Quiet 80%, Normal 79%, Loud 82%.

VI. CONCLUSION

We demonstrate the ability to localize people using only the subtle sounds they incidentally produce as they move. We present the Robot Kidnapper dataset and the resulting person detection models that can be implemented on robots to track a person as they move quietly. Our work opens up an avenue of exploration for how robots can learn human awareness with only passive audio sensing and without nearby humans needing to intentionally produce additional sounds.

Limitations: Our dataset only contains instances of a single person in a room up to a maximum distance of 6m. We also have not tested our method on different types of microphones or robot setups. Additionally, we are unable to localize a person if they are standing completely still.

References

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[4] Timm Linder, Narunas Vaskevicius, Robert Schirmer, and Kai Oliver Arras. Cross-modal analysis of human detection for robotics: An industrial case study. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 971–978, 2021.

[5] Dan Jia, Alexander Hermans, and Bastian Leibe. Domain and modality gaps for lidar-based person detection on mobile robots. *arXiv e-prints*, pages arXiv–2106, 2021.

[6] Dan Jia, Alexander Hermans, and Bastian Leibe. Dr-spaam: A spatial-attention and auto-regressive model for person detection in 2d range data. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10270–10277. IEEE, 2020.

[7] Zhi Yan, Li Sun, Tom Duckctr, and Nicola Bellotto. Multisensor online transfer learning for 3d lidar-based human detection with a mobile robot. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7635–7640, 2018.

[8] Angus Fung, Beno Benhabib, and Goldie Nejat. Robots autonomously detecting people: A multimodal deep contrastive learning method robust to intraclass variations. *IEEE Robotics and Automation Letters*, 2023.

[9] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv:1802.02611*, 2018.

[11] Timm Linder, Dennis Grießer, Narunas Vaskevicius, and Kai Oliver Arras. Towards accurate 3d person detection and localization from rgb-d in cluttered environments. 2018.

[12] T. Linder, K. Y. Pfeiffer, N. Vaskevicius, R. Schirmer, and K. O. Arras. Accurate detection and 3D localization of humans using a novel YOLO-based RGB-D fusion approach and synthetic training data. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[13] Lucas Beyer, Alexander Hermans, Timm Linder, Kai Arras, and Bastian Leibe. Deep person detection in 2d range data. *IEEE Robotics and Automation Letters*, PP, 04 2018.

[14] Dan Jia, Alexander Hermans, and Bastian Leibe. 2D vs. 3D LiDAR-based Person Detection on Mobile Robots. *arXiv e-prints*, page arXiv:2106.11239, June 2021.

[15] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L. Waslander. Joint 3d proposal generation and object detection from view aggregation. page 1–8. IEEE Press, 2018.

[16] Timm Linder, Stefan Breuers, Bastian Leibe, and Kai O. Arras. On multi-modal people tracking from mobile platforms in very crowded and dynamic environments. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5512–5519, 2016.

[17] Gaurav Singh, Paul Ghanem, and Taskin Padir. Sporadic audio-visual embodied assistive robot navigation for human tracking. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '23, page 99–105, New York, NY, USA, 2023. Association for Computing Machinery.

[18] Yuki Yamazaki, Chinthaka Premachandra, and Chamika Janith Perea. Audio-processing-based human detection at disaster sites with unmanned aerial vehicle. *IEEE Access*, 8:101398–101405, 2020.

[19] Ilaria Ciuffreda, Gianmarco Battista, Sara Casaccia, and Gian Marco Revel. People detection measurement setup based on a doa approach implemented on a sensorised social robot. *Measurement: Sensors*, 25:100649, 2023.

[20] Marco Sewtz, Tim Bodenmüller, and Rudolph Triebel. Robust music-based sound source localization in reverberant and echoic environments. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2474–2480, 2020.

[21] Alban Portello, Patrick Danès, and Sylvain Argentieri. Acoustic models and kalman filtering strategies for active binaural sound localization. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 137–142, 2011.

[22] D. Bechler, M.S. Schlosser, and K. Kroschel. System for robust 3d speaker tracking using microphone array measurements. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, volume 3, pages 2117–2122 vol.3, 2004.

[23] Zhanbo Shi, Lin Zhang, and Dongqing Wang. Audio-visual sound source localization and tracking based on mobile robot for the cocktail party problem. *Applied Sciences*, 13(10), 2023.

[24] Deepak Gala, Nathan Lindsay, and Liang Sun. Realtime active sound source localization for unmanned ground robots using a self-rotational bi-microphone array. *Journal of Intelligent & Robotic Systems*, 95:935 – 954, 2018.

[25] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[26] Ziyang Chen, David F Fouhey, and Andrew Owens. Sound localization by self-supervised time delay estimation. *European Conference on Computer Vision (ECCV)*, 2022.

[27] Abhinav Tripathi, Mohd. Ammar Khan, Akash Pandey, Pankaj Yadav, and Amit Kumar Sharma. Human following robot using ultrasonic sensor. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pages 764–770, 2021.

[28] Jesper Haahr Christensen, Sascha Hornauer, and Stella X. Yu. Batvision: Learning to see 3d spatial layout with two ears. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1581–1587, 2019.

[29] Mason Wang, Samuel Clarke, Jui-Hsien Wang, Ruohan Gao, and Jiajun Wu. Soundcam: A dataset for finding humans using room acoustics. In *Advances in Neural Informaion Processing Systems*, 2023.

[30] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[31] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022.

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[33] Cristiano Premebida, Joao Carreira, Jorge Batista, and Urbano Nunes. Pedestrian detection combining rgb and dense lidar data. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4112–4117. IEEE, 2014.

[34] Pedro J Navarro, Carlos Fernandez, Raul Borraz, and Diego Alonso. A machine learning approach to pedestrian detection for autonomous vehicles using high-definition 3d range data. *Sensors*, 17(1):18, 2016.

[35] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[36] Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1257–1265, 2020.

[37] Jisoo Park, Jingdao Chen, Yong K Cho, Dae Y Kang, and Byung J Son. Cnn-based person detection using infrared images for night-time intrusion warning systems. *Sensors*, 20(1):34, 2019.

[38] Thi Thi Zin, Hideya Takahashi, Takashi Toriu, and Hiromitsu Hama. Fusion of infrared and visible images for robust person detection. *Image fusion*, pages 239–264, 2011.

[39] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey

and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[40] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6036–6046, 2018.

[41] Samarth Brahmbhatt, Heni Ben Amor, and Henrik Christensen. Occlusion-aware object localization, segmentation and pose estimation. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 80.1–80.13. BMVA Press, September 2015.

[42] Nicola Maria Ceriani, Giovanni Buizza Avanzini, Andrea Maria Zanchettin, Luca Bascetta, and Paolo Rocco. Optimal placement of spots in distributed proximity sensors for safe human-robot interaction. In *2013 IEEE International Conference on Robotics and Automation*, pages 5858–5863. IEEE, 2013.

[43] Christian Zieger, Alessio Brutti, and Piergiorgio Svaizer. Acoustic based surveillance system for intrusion detection. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 314–319. IEEE, 2009.

[44] Selina Chu, Shrikanth Narayanan, C-C Jay Kuo, and Maja J Mataric. Where am i? scene recognition for mobile robots using audio features. In *2006 IEEE International conference on multimedia and expo*, pages 885–888. IEEE, 2006.

[45] Jonas Hornstein, Manuel Lopes, José Santos-Victor, and Francisco Lacerda. Sound localization for humanoid robots-building audio-motor maps based on the hrtf. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1170–1176. IEEE, 2006.

[46] Hongzhuo Liang, Shuang Li, Xiaojian Ma, Norman Hendrich, Timo Gerkmann, Fuchun Sun, and Jianwei Zhang. Making sense of audio vibration for liquid height estimation in robotic pouring. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5333–5339. IEEE, 2019.

[47] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. In *5th Annual Conference on Robot Learning*, 2021.

[48] J-M Valin, François Michaud, Jean Rouat, and Dominic Létourneau. Robust sound source localization using a microphone array on a mobile robot. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, volume 2, pages 1228–1233. IEEE, 2003.

[49] Justin Wilson and Ming C Lin. Avot: Audio-visual object tracking of multiple objects for robotics. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10045–10051. IEEE, 2020.

[50] Yoko Sasaki, Ryo Tanabe, and Hiroshi Takernura. Online spatial sound perception using microphone array on mobile robot. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2478–2484. IEEE, 2018.

[51] Arun Balajee Vasudevan, Dengxin Dai, , and Luc Van Gool. Semantic object prediction and spatial sound super-resolution with binaural sounds. *ECCV*, 2020.

[52] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.