Effects of Robot Sound on Auditory Localization in Human-Robot Collaboration

Elizabeth Cha[1], Naomi T Fitter[1], Yunkyung Kim[2], Terrence Fong[2], Maja J Matarič[1]

[1]Department of Computer Science, University of Southern California, Los Angeles, CA

[2]Intelligent Robotics Group, NASA Ames Research Center, Mountain View, CA

ABSTRACT

Auditory cues facilitate situational awareness by enabling humans to infer what is happening in the nearby environment. Unlike humans, many robots do not continuously produce perceivable state-expressive sounds. In this work, we propose the use of iconic auditory signals that mimic the sounds produced by a robot's operations. In contrast to artificial sounds (e.g., beeps and whistles), these signals are primarily functional, providing information about the robot's actions and state. We analyze the effects of two variations of robot sound, tonal and broadband, on auditory localization during a human-robot collaboration task. Results from 24 participants show that both signals significantly improve auditory localization, but the broadband variation is preferred by participants. We then present a computational formulation for auditory signaling and apply it to the problem of auditory localization using a human-subjects data collection with 18 participants to learn optimal signaling policies.

CCS CONCEPTS

• **Human-centered computing** → **Auditory feedback**; *Empirical studies in HCI*; • **Computer systems organization** → **Robotics**; • **Computing methodologies** → Robotic planning;

KEYWORDS

human-robot interaction, sound, auditory localization, collaboration, coordination, nonverbal communication

1 INTRODUCTION

Many robots lack nonverbal cues that enable humans to infer information about the robot's cognitive and physical state [3, 10, 20, 23, 30]. Auditory cues provide a wide range of contextual information that promote awareness of a person's surroundings [9, 24]. Humans naturally and often subconsciously produce auditory signals that reveal information about their current physical state, such as breathing sounds or footsteps [24, 35]. Since many robots do not produce these state-expressive and perceivable sounds, it can be difficult for a human to use auditory cues to localize a robot in their surroundings, as illustrated in Fig.1.

As human knowledge of a robot's location is key to both safety and collaboration, the goal of this work is to explore robot sounds that facilitate auditory robot localization by co-located humans. A straightforward solution is for the robot to employ a loud, distinctive sound to indicate its presence. However, more noticeable auditory signals are also more likely to annoy or distract people [7, 31]. Instead, the design of the robot's auditory cues should take into account the preferences of the human collaborator.

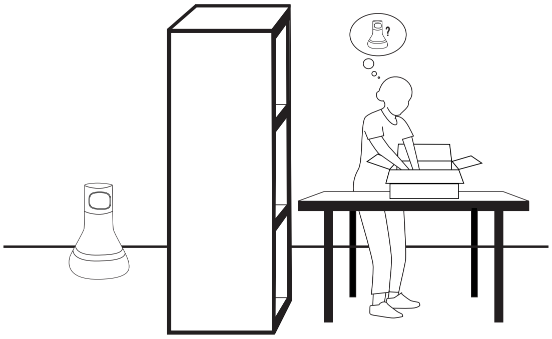


Figure 1: This work explores the effects of different auditory signals on a human's ability to localize a co-located robot.

Prior work has demonstrated that artificial sounds (e.g., beeps and chirps), like those used by the fictional robots R2D2 and WALL-E, can convey social properties, such as affect and politeness [28]. As these sounds more closely mimic linguistic properties, however, they are less suited for conveying concrete information about the robot's state [29]. Section 2 provides background on related auditory signaling and localization research.

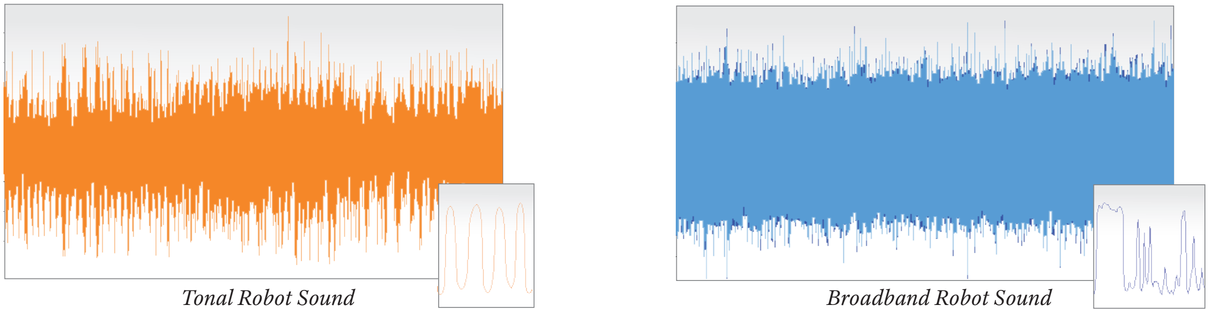
In Section 3, we propose the use of robot sounds that act as *auditory icons* designed to mimic the noise normally produced by a robot's operations. These sounds can be modulated to enhance perception, augmenting the robot's natural cues. Since these sounds more closely match humans' expectations, they are also easier to interpret than melodies or synthetic cues.

Towards the goal of enabling auditory localization in a user-acceptable manner, we present an initial user study with 24 participants in Section 4. We compare two variations of robot auditory signal, broadband and tonal, in a human-robot collaboration (HRC) task. *Broadband sounds* contain a larger range of frequencies, and are thus less annoying than traditional tonal indicators or alarms. *Tonal sounds*, however, are typically more distinctive and noticeable.

In Section 5, we present the results of the user study which reveal that both variations of auditory signal enabled more accurate and faster localization, positively affecting the HRC. We also found that although participants preferred the broadband sound, neither signal was significantly easier or faster to localize. Overall, these findings support the use of iconic robot sounds for auditory localization and in particular, broadband sounds for reduced user annoyance.

To build on the user study findings and broaden the applicability of this work, we also propose a model for planning auditory signals in Section 6. This model considers information about the scenario to determine optimal auditory signals. The goal of the model is to

HRI '18, March 5–8, 2018, Chicago, IL, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-4953-6/18/03…$15.00
https://doi.org/10.1145/3171221.3171285



Tonal Robot Sound

Broadband Robot Sound

Figure 2: Two auditory signal variations, tonal and broadband, used in the initial user study; the tonal has a regular, distinctive frequency and the broadband is a mixture of many frequencies. Both were generated from recordings of the Turtlebot's sound while driving.

enable people to quickly localize robots in their environment with minimal annoyance, encoded as sound duration and intensity. We performed a human-subjects data collection with 18 participants to learn optimal signaling policies offline.

In Section 7, we discuss the overall results of this work, its limitations, and its contributions to the field of HRI.

2 BACKGROUND

Communicating information about a robot's state is a growing area of HRI research. Recent works have proposed various methods for expressing intent through robot motion [11, 18, 32] and conveying information about the robot's task through light [2, 7]. As many robots are functional by design, these works often draw inspiration from design principles for appliances, cars, and other devices [7, 33].

The primary challenge for nonverbal robot communication is designing signals that are intuitive and easy to interpret [7]. Although training or prolonged exposure can help humans understand many everyday nonverbal signals from electronic devices [13, 16], robots often require a wider variety of signals than most people can easily learn. Signals that are analogous to cues people already understand may help to convey diverse nonverbal robot communication.

This approach is especially important when using communication modalities that are less precise, such as sound. While vision is traditionally preferred in applications that require a high level of perceptual discrimination [4, 8], we envision robots to to function in many scenarios and environments where the visual field is limited, blocked, or noisy [9, 19]. In these situations, auditory cues can augment a human's knowledge of a co-located robot's state, including its presence and location in the environment [9].

Knowledge of the robot's location enables coordination and prevents collisions [1, 12]. This is crucial as many robots are limited in their ability to quickly detect and avoid dynamic obstacles, putting the burden of safety on the human. Humans and animals often utilize sounds to estimate the position of others in a process called auditory localization [21], which is relatively unexplored for HRI. Past approaches have instead relied on visual cues such as motion or light to increase awareness of a robot [2, 7].

Auditory localization relies on several perceptual phenomena including interaural time difference (ITD), the time difference between the arrival of the same sound at each ear, and interaural intensity difference (IID), the difference in intensity (i.e., loudness) of the same sound between ears [5, 14]. Although auditory localization is quite complex, we simplify the problem by only looking at relative localization of a sound source in two dimensions.

To inform robot sound design, we also explored two variations of sound. Tonal sounds are often used for alarms and alerts because they are difficult to ignore; this also makes them well suited for less frequent emergency scenarios where noticeability is key [25, 27]. Broadband sounds contain a larger range of frequencies, making the signal less distinctive and more similar to ambient noise [14, 26] and have been found to be easier to localize than pure tones [21].

3 SOUND DESIGN

To explore how robots can employ auditory signals to enable localization by humans, we generated and compared a tonal sound with a discrete tone and a broadband sound with many frequencies. Both variations of auditory signal were designed to convey that the robot is in motion. Samples of the sounds can be found at [6].

3.1 Auditory Icon

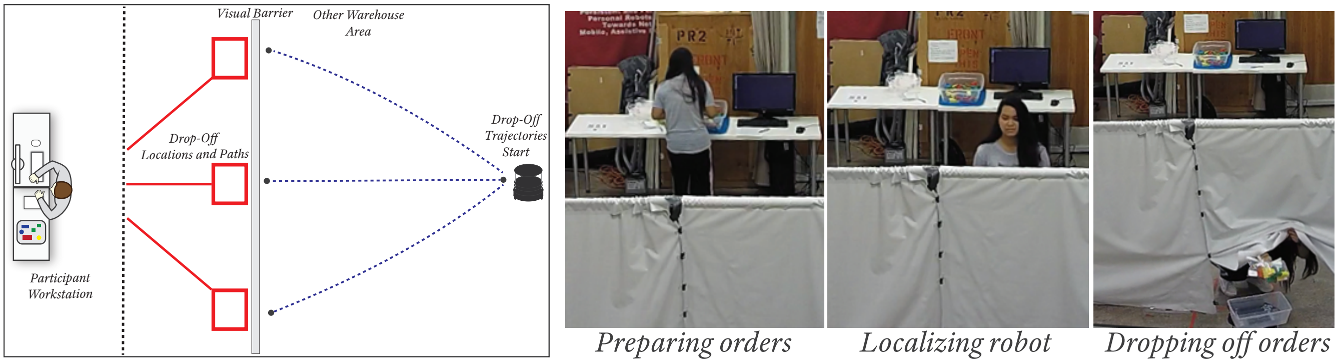
Auditory icons are sounds that are designed to act as analogies to typical sounds or noises found in everyday life [17]. Instead of sounds with learned meanings, these sounds are more iconic, matching people's expectations and experiences. For instance, to indicate that a robot is powered on, it can mimic sounds that other machinery or people emit (e.g., computer humming and breathing).

Since robots take many different forms, they produce a wide range of noises with varying levels of noticeability. Recent research has sought to discover what properties these sounds naturally convey to human listeners [22, 34]. Using this knowledge, auditory cues with specific meanings (e.g., motion) can be modulated to alter humans' mental perceptions of the robot.

The use of such icons can facilitate standardization of auditory signals across robot platforms and reduces the number of cues humans are required to learn. As the proposed auditory icons are expressive of the robot's actual state, these signals are more intuitive and easier to learn compared to melodies or artificial sounds. The use of auditory icons reduces the need for prior exposure or training.

3.2 Tonal Robot Sound

To create a tonal robot sound, we first recorded the sound produced by the motion of the Turtlebot 2, the first mobile robot platform utilized in this work. The goal for the tonal signal was to create a machine-like sound that clearly resembled the periodic noise of the robot's motors while in motion. Then, we removed the background noise and other unwanted noise components from the raw auditory signal. This made the sound more distinctive and clearer than the original recording. To make the sound more noticeable, we also altered the pitch to be about 10% higher, as humans tend to be more sensitive to higher-frequency noises within a certain range (Fletcher-Munson curves [15]). The end result was a regular servo-like sound at a frequency of 700 Hz (Fig.2, left).



Preparing orders Localizing robot Dropping off orders

Figure 3: An overview of the experimental setup. The diagram on the left shows the setup schematic, and the photographs on the right show the interaction process in the warehouse area.

3.3 Broadband Robot Sound

Although broadband sounds do not have a distinctive pitch, we wanted the sound to remain reminiscent of the robot’s motor sounds. The broadband signals consisted of the tonal robot sound with added Brownian noise, generated by integrating white noise with equal intensity throughout the frequencies found within the human hearing range (20 Hz-20 kHz). Compared to white noise, Brownian noise is considered less harsh as it has more energy at lower frequencies, giving it a softer quality. We chose Brownian noise for its resemblance to the deeper, rumbling sounds typical of machinery. The resulting sound (Fig.2, right) still contained a distinctive "wheel turning" component but had no clear tone or frequency.

4 EXPERIMENTAL DESIGN

To explore the effects of iconic robot sounds on human-robot collaboration, we conducted a within-subjects user study in which participants collaborated with the Turtlebot 2 on a physical task.

4.1 Hypotheses

Since the sounds produced by many robots are difficult to perceive against ambient noise, we anticipated the addition of either type of auditory signal will improve participants’ ability to localize the robot. To evaluate this, we introduced a baseline sound condition in which no additional auditory signal is present. We also expected that one sound type will be preferred overall.

H1: Objective Collaboration Metrics. *Auditory condition (baseline, tonal, broadband) will significantly affect participants’ accuracy and time when inferring the robot’s location, with broadband sound being the best.*

H2: Perception of the Sounds. *The tonal and broadband signals will negatively affect participants’ perceptions of annoyance but will positively affect noticeability and localizability of the robot. Participants will perceive the broadband sound to be the easiest to localize and the tonal sound to be the most noticeable and annoying.*

H3: Sound Type Preference. *Participants will prefer the broadband robot sound to the tonal robot sound for use in a human-robot collaborative task.*

4.2 Collaborative Task

Designing a collaborative task to analyze the effects of the sound type presented several challenges. To isolate the effects of auditory signals on localization, visual cues needed to be blocked or minimized. Since user performance needed to be compared across the sound conditions, the task also needed to be repeatable and consistent; positioning was especially important as variations in distance or position can significantly affect sound perception. The participants’ actions also needed to depend on their ability to localize the robot to provide observable, objective metrics. Lastly, the task had to be realistic to create an authentic experience for participants and motivate their performance. This authentic framing also provides insight into how the results of this work can be applied to real applications and environments.

To satisfy the task criteria, we chose a collaborative packing activity in a mock warehouse environment. The participant and robot worked together to package orders and distribute them across the environment. The participant’s role is to find the individual items (building blocks) for each order, place them in a bag with the corresponding order number, and hand off the package to the robot. The robot periodically retrieved the participant’s completed orders and transported them across the “warehouse.”

For this experiment, we used the Turtlebot 2 mobile robot base with two additions: 1) a USB speaker mounted on top of the robot for playing the auditory signals and 2) a plastic bin mounted behind the speaker for holding completed orders.

Fig.3 shows a schematic of the task. The participant is on one side of the room separated from the robot by a 5 foot tall white curtain. The participant continuously packages orders until a message on the monitor asks whether they are ready for the robot to pickup the orders. After the participant confirms, the robot moves to one of the three drop-off areas behind the white curtain. The participant collects their completed orders, chooses the drop-off area (left, right, or middle) they believe the robot is located at, and moves to the area using the corresponding path marked by red tape. After the participant walks a certain distance (marked by gray dashed line) past the workstation, they are not allowed to change paths. This enables us to clearly tell which drop-off area the participant has chosen. The participant pulls up the curtain, places the orders in the robot’s bin, and returns to their workstation.

To motivate their performance, participants were told they would receive a bonus based on their accuracy and speed, such that an incorrect drop-off location would be penalized and a faster drop-off time would be rewarded. Hence, participants aimed to choose the correct drop-off location as quickly as possible. We also told participants that completing the packaging of orders slightly affected the bonus to prevent them from solely focusing on the robot.

Objective Collaboration Measures

	Accuracy	Avg. Time (sec)
Baseline	36.8% (53/144)	11.90 (*SD*=8.47)
Tonal	91.7% (132/144)	7.56 (*SD*=5.36)
Broadband	91.0% (131/144)	6.93 (*SD*=4.82)
	F(2,432)=108.67 *p*<0.001	F(2,432)=25.69 *p*<0.001

▨ Baseline ▨ Tonal ▨ Broadband

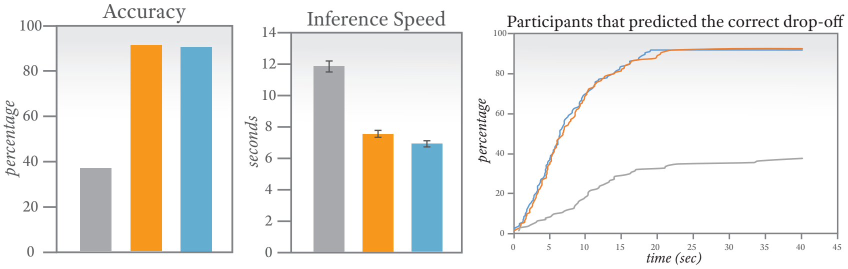


Figure 4: Objective collaboration metrics: the drop-off prediction accuracy and the decision speed. Left and middle show the means and results of repeated-measure ANOVAs. The right shows the cumulative percent of correct drop-offs chosen by time.

Visual Barrier: In real world scenarios, humans utilize both visual and auditory cues for localization, but we used a visual barrier to prevent participants from fixating on the robot due to its novelty, the design of the experiment, and the bonus. This approach also allowed us to isolate the effects of the auditory cues. Once participants reached their chosen drop-off location, they were allowed to look over the barrier to check whether their prediction was correct.

Robot Trajectories: In order to know when to start listening for the robot, participants triggered the execution of the robot's pick-up trajectory. The robot utilized three straight line trajectories starting at the middle of the room, opposite the participant to get to the drop-off areas. The robot moved at 0.25 meters per second to give participants enough time to observe the auditory signal before making a prediction.

Environmental Noise: We also altered the ambient noise levels of the environment to create an effect consistent with a real factory or warehouse. Industrial settings typically have sound measurements ranging from 75 to 90 dBA. We utilized a continuous sound track of industrial noise for the duration of the experiment. Using a sound level meter, we adjusted the sound of the environment to be about 70 dBA. We chose this level because the experimental area was smaller and more enclosed than a real warehouse.

4.3 Procedure

As participants entered the experiment room, they were given a written overview of the study with a picture of the robot. After obtaining informed consent, the experimenter explained the collaborative task in detail. Participants were told that the goal of the study was to explore methods to better coordinate human and robot actions using sound. Participants were also told they could earn a 67% bonus (added to their compensation) based on the performance metrics described in the previous Section. After finishing the task, participants completed a post-study survey.

4.4 Manipulated Variables

We manipulated a single variable, *auditory signal type*, the sound that the robot plays while it is moving to be *baseline* (i.e., absent), *tonal*, or *broadband*. The baseline sound condition has no added auditory signal and no amplification; it represents a more typical use case in which a robot moves about the environment emitting only the noises it naturally produces.

A sound level meter was used to make the tonal and broadband auditory signals equal in sound level intensity (i.e., volume). Due to the robot's short height and the presence of a curtain, we had to amplify the signals to overcome the physical interference. No sound was played while the robot was stationary.

4.5 Participants

A total of 24 participants (15 males, 9 females, ages 18-35, *M*=23.83, *SD*=4.25) were recruited from the local community. Four participants reported having experience working with or using robots.

The experiment used a within-subjects design as it enabled participants to compare the three different auditory signal conditions. Each auditory condition was used for six drop-offs, equally divided among the three locations. The order of drop-offs and the conditions were fully counterbalanced to control for ordering effects.

Participants were told that there were three different types of sounds that the robot produced. They were also informed of and given a short break when a transition occurred between the conditions to avoid confusion from a sudden sound change. To prevent participants from continuously listening for the sounds, the robot drove around the warehouse between drop-offs. This also enabled participants to familiarize themselves with each auditory condition.

4.6 Dependent Measures

Both objective and subjective measures were used to evaluate the effects of each auditory condition on participants' ability to localize the robot during a HRC task. The objective measures include the *decision time* and the *accuracy* for choosing drop-off locations. The decision time is the amount of time it takes for the participant to choose a drop-off location after prompting the robot to start its trajectory. Each drop-off was scored as either 0 or 1 in accuracy.

The subjective measures consisted of 5-point Likert scale ratings for each auditory condition's *noticeability*, *localizability*, and *annoyance.* Additionally, participants were forced to choose an auditory condition for each of these descriptors: most noticeable, easiest to localize, most annoying, and most preferred to work with. As a manipulation check, participants also described each of the auditory conditions.

5 RESULTS

The experimental task was divided into three sessions, one for each auditory condition. Each session consisted of six drop-offs. This led to a total of 432 interactions or trials.

H1- Objective Collaboration Metrics: Our first hypothesis stated that the auditory condition (baseline, tonal, broadband) will significantly affect the objective collaboration metrics. We also predicted that the broadband signal will perform the best.

To test this hypothesis, we performed repeated-measures ANOVAs on participants' prediction accuracy and decision time (i.e., inference speed). We found that sound type significantly affected both accuracy ($F(2, 432) = 108.67, p < 0.001$, Fig.4) and inference speed ($F(2, 432) = 25.69, p < 0.001$, Fig.4). However, a post hoc analysis

	Subjective Collaboration Measures		
	Noticeability	Localizability	Annoyance
Baseline	1.125 (*SD*=0.34)	1.125 (*SD*=0.34)	1.083 (*SD*=0.28)
Tonal	4.750 (*SD*=0.53)	4.083 (*SD*=0.72)	3.750 (*SD*=0.85)
Broadband	4.292 (*SD*=0.69)	4.375 (*SD*=0.65)	2.875 (*SD*=0.99)
	F(2,24)=321.21 *p*<0.001	*F*(2,24)=222.32 *p*<0.001	*F*(2,24)=74.71 *p*<0.001

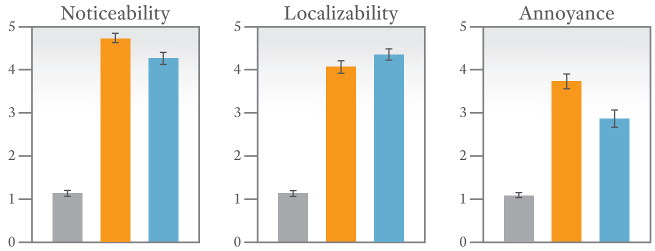


Figure 5: Average ratings on a 5-point Likert scale for the subjective collaboration metrics: noticeability, localizability, and annoyance of each sound condition.

using Tukey HSD showed that the tonal and broadband conditions were not significantly different for either metric.

Overall, participants had a lower average inference time for the broadband condition while the accuracy between the two conditions was about the same. Several participants commented that the tonal signal was easier to localize from the start of the session, whereas the broadband signal took some time to get used to.

Analysis of the video recordings also showed that several participants hesitated in the initial drop-off trials for the broadband condition. They moved their bodies back and forth before slowly walking to the drop-off location. This suggests that the broadband sound is less intuitive and may require familiarization time.

H2- Perceptions of the Sounds: Our second hypothesis stated that addition of auditory signals (i.e., tonal and broadband conditions) will negatively affect participants' ratings of annoyance, but positively affect ratings of robot noticeability and localizability. We also predicted that participants will perceive the broadband signal as the easiest to localize and the tonal signal to be the most noticeable and annoying.

We performed repeated-measures ANOVAs on participants' ratings of all three subjective collaboration metrics. As predicted, we found that auditory signal condition significantly affects all three measures (Fig.5): noticeability ($F(2, 24) = 321.21, p < 0.001$), localizability ($F(2, 24) = 222.32, p < 0.001$), and annoyance ($F(2, 24) = 74.71, p < 0.001$).

Participants rated the tonal signal as the most annoying and noticeable, while the broadband signal was rated the most localizable. A post hoc analysis with Tukey HSD confirmed that each auditory condition was significantly different for noticeability and annoyance ratings. As with our objective collaboration metrics, we found that the added auditory signal conditions were rated as significantly more localizable than the baseline condition, but that the tonal and broadband conditions were not rated significantly differently from each other.

In a set of fixed-choice questions, we also asked participants to choose the auditory conditions they thought was the most noticeable, easiest to localize, and most annoying. 88% of participants chose the tonal condition as the most noticeable while 83% of participants also chose this signal as the most annoying. In contrast, 67% of participants chose the broadband robot as the easiest to localize, confirming the mixed results between the two added auditory signal conditions.

Participants commented that the tonal signal was the most annoying and noticeable due to its high pitch. Several participants also mentioned that the tonal stood out from the ambient sounds, compared to the broadband signal, which sounded "lower" and more like the factory noise.

H3- Sound Type Preference: Our final hypothesis predicted that participants will prefer the broadband auditory condition to the tonal auditory condition for use in a human-robot collaboration task. The post-study survey asked participants to choose the robot that they would prefer to work with and explain their selection.

79% of participants prefer to work with the robot with the broadband auditory condition while 21% of participants prefer the robot with the tonal condition. Participants' comments reveal that they felt that they were able to find both robots in the environment, but that the tonal sound condition was more annoying.

Participants also commented that the higher pitch of the tonal signal was more distracting and difficult to ignore. A handful of participants also stated that they would be unable to work for a long period of time in the presence of the tonal condition robot. More than one participant commented that they would, "lose it" if they had to listen to the tonal sound for a full workday. Another participant summarized the experiment by saying, "the first sound (baseline) would cause me to lose my job in the first week and the second (tonal) would make me quit in the first week."

6 AUDITORY SIGNALING MODEL

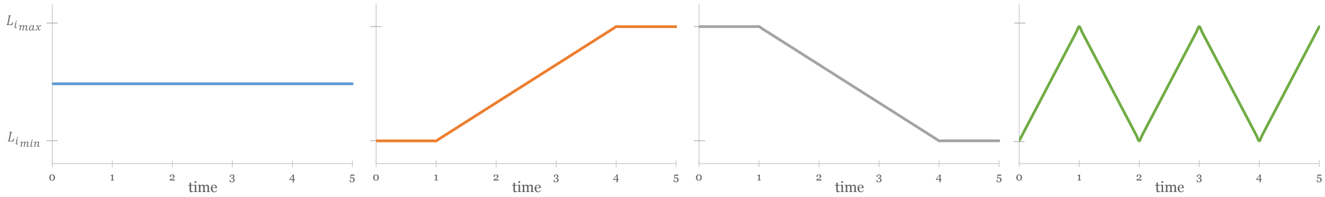
The previously described user study explored the space of auditory icons to help humans localize robots during HRC; the results of this investigation support the use of broadband signals for user-accepted auditory robot localization. To isolate the effects of auditory cues, this user study employed a visual barrier to separate the robot and participants, but humans typically utilize a mixture of auditory and visual cues for localization.

To apply the initial user study results in a more realistic setting, we propose a general auditory signaling model that incorporates visual cues and information about the environment when deciding the robot's auditory signaling policy. We apply this model to auditory localization to learn optimal auditory signal policies across different world states.

This model enables the robot to account for uncertainty in the world, including human perception, when planning its signals. Since past work in auditory localization has primarily taken place in controlled laboratory settings, the goal of this work is to apply the results of the initial user study to the dynamic settings robots are expected to operate in. This generalization enables us to create more intelligent robot behaviors and explore more dynamic auditory signals than the static sounds employed in the initial user study.

6.1 Model Formulation

In this work, we formulate the problem of auditory signaling as a Markov Decision Process (MDP). MDPs can be described as a tuple $\{S, A, T, R\}$, where:



(1) Steady (2) Increasing (3) Decreasing (4) Triangle

Figure 6: The auditory signaling policies utilized in the experiment. The sound intensity, L_i, is varied over time where $t = 0$ is the time when a person is cued to find the Ava robot in our data collection. The minimum and maximum values of L_i depend on the ambient sound level e_s.

- S is a finite set of world states modeling different environment configurations.
- A is a finite set of actions that the robot can take. In this model, the robot's actions are the set of auditory signals the robot can employ.
- $T : S \times A \rightarrow \Pi(S)$ is the state-transition function that gives a probability distribution over world states for each state and action. It is typically represented as $P_a(s, s')$ or the probability that action a in state s will lead to state s'. The transition function models the variability in human response to an auditory signal.
- $R : S \times A \rightarrow \mathbb{R}$ is the reward function, giving the expected immediate reward gained by taking each action in each state. It is represented by $R(s, a)$ or the expected reward for taking action a in state s.

The policy π of the robot is the assignment of an auditory signal action $\pi(s)$ at every state s. The optimal policy $\pi^*(s)$ is the auditory signaling policy that maximizes the reward. We can also incorporate a discount factor γ to balance immediate and long term rewards.

6.2 Auditory Localization MDP

We apply the MDP model above to auditory signaling to enable localization of the robot by co-located humans. In this scenario, we assume that for each state, there is a set of signaling actions and an associated reward distribution that is stationary. Hence, we formulate the problem as a set of one-state MDPs and learn the best policy (i.e., signaling action) for each MDP or world state.

States: In the previous user study, we limited participants' sensing capabilities by removing visual cues from the environment. To more accurately represent how humans localize other agents in the real world, we assume that visual cues will also be utilized and incorporate this knowledge of the environment into the model. The resulting world state s is a tuple $\{e_s, v, d\}$ consisting of the ambient sound level of the environment e_s, the visibility of the robot from the human observer's perspective v, and the distance between the robot and human d. We discretized each of these variables in our experiment as follows.

- $e_s = \{1, 2\}$- We utilized the same sound track from our first experiment at two different sound levels: 1) 55 dbA (similar levels to an office or restaurant) and 2) 70 dbA (similar to an industrial environment).
- $v = \{1, 2, 3, 4\}$- Visibility of the robot takes into account whether the robot is in the human's viewing range (approximately 114 degrees) and whether the view of the robot is obstructed by objects in the environment. This results in four levels of visibility (lowest to highest): 1) obstructed and out of view, 2) obstructed and in view, 3) not obstructed and out of view, and 4) not obstructed and in view.
- $d = \{1, 2\}$- Due to the limited size of the experimental space (17 feet by 19.5 feet), the distance d was varied to be either 1) close or 2) far.

Actions: We used the same auditory signal for all signaling policies $\pi(s)$. The signal is a broadband sound recorded from the iRobot AVA mobile base, the robot utilized in the experiment. The recording overall resembles the sounds produced by operation of a desktop computer. We created four auditory signaling policies by varying the sound level intensity over time, as shown in Fig.6.

The first policy (Fig.6 (1)) keeps the sound level consistent. The second policy (Fig.6 (2)) linearly increases the sound level over time until it reaches a maximum value. The third policy (Fig.6 (3)) linearly decreases the sound level over time until it reaches a minimum value. Policies (2) and (3) both maintain a constant sound intensity level for the first second to give participants time to notice the stimuli in a later introduced data collection task. After reaching the minimum or maximum sound intensity level at $t = 4$, the sound intensity level for these two policies stays constant for the remainder of the interaction. The last policy (Fig.6 (4)) linearly increases and decreases the sound intensity to reach these maximum and minimum levels at a regular frequency.

The minimum and maximum sound intensity levels ($L_{i\,min}$ and $L_{i\,max}$) were chosen based on the ambient sound level of the environment. For $e_s = 1$, $L_{i\,min} = 65$ dbA and $L_{i\,max} = 75$ dbA at the sound source. For $e_s = 2$, $L_{i\,min} = 80$ dbA and $L_{i\,max} = 90$ dbA at the sound source.

We chose to explore a narrowly focused set of policies to support our goal of investigating an initial application of the proposed auditory signaling model. Moreover, for the task of auditory localization, the robot is expressing only one internal state feature, making it logical to utilize only one auditory signal. Variation of the sound level intensity enables us to alter the salience or noticeability of the signal, which may be more important in cases where the human observer cannot utilize visual cues.

Reward: A key insight from our first study was the need for robot signals that minimize human annoyance. Therefore, our reward function aims to balance the human's ability to localize the robot with the signal's annoyance. In this work, we equate annoyance to the sound level of the robot's auditory signal. We formulate the reward as

$$R = \frac{1}{t_{resp_{norm}}} \times \frac{1}{\int_0^{t_{resp}} L_i(t)dt} \tag{1}$$

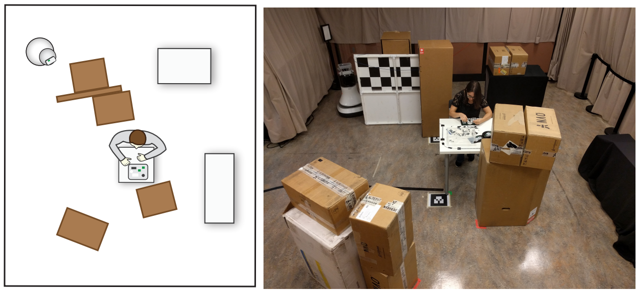


Figure 7: The diagram on the left shows the setup schematic, and the photograph on the right shows the actual data collection space.

where t_{resp} is the participants response time, $t_{resp_{norm}}$ is the response time normalized by distance to the robot, and $L_i(t)$ is the sound level intensity of the auditory signal over time.

The reward values a lower response time, as shown by the first term in (1). The second term penalizes using higher sound level intensities, especially for longer periods of time. Hence, auditory signals that cause the human to respond quickly or that do not have high intensity levels may yield a better reward than signals with a constant intensity level, such as those employed in Section 4.

6.3 Auditory Localization Data Collection

We applied the proposed model to auditory localization using a human-subjects data collection to directly learn the optimal signaling policy for a subset of world states. Auditory localization is primarily important for scenarios when a robot and human are in relatively close proximity, so we discretized the most relevant world state space and directly learned a reward distribution for each world state.

Data Collection Task: Throughout the data collection task, a human participant sat at a workstation constructing a LEGO structure, as shown in Fig.7. The participant received a bin of LEGO pieces and assembly instructions. Participants were instructed to work on the structure at all times unless provided with a light-based cue to interact with the robot.

The robot is introduced as a way to provide the participant with additional parts for their LEGO structure. When the LED lights surrounding the participant's workstation turn green, the robot begins its auditory signaling policy. This is a prompt for the participant to leave the workstation, go to the robot, retrieve a LEGO piece, and return to their workstation task. Both the LEDs and auditory signal are turned off when the participant returns to the workstation.

Similar to the order packaging task from the initial user study, the structure building task was designed to distract the participant and force them to focus on something other than the robot. The robot also randomly moves around the room between interactions to familiarize participants with the auditory and visual cues provided by its motion. Ambient environment sounds were provided by four speakers located around the perimeter of the room.

Robot: The robot used in the data collection was an iRobot AVA holonomic mobile base. The robot was equipped with a cylindrical speaker and a bin for holding LEGO blocks.

Measures: For each interaction, we measured the time for the participant to reach the robot (t_{resp}) and this response time normalized by the starting distance between the robot and human ($t_{resp_{norm}}$).

Participants: A total of 16 participants (11 males, 5 females, ages 19-31, M=27.5, SD=3.92) were recruited from the local community for the data collection. We chose to expose each participant to only one of the ambient sound levels e_s, limiting the total number of interactions and avoiding fatigue. We exposed each participant to the remaining 8 world states (4 visibility levels and 2 distances) and 4 auditory signaling policies, for a total of 32 interactions.

6.4 Results

The goal of this data collection was twofold: to apply the proposed signaling model to a real problem and to gather insights into auditory localization of a robot by co-located humans. Since few works have explored auditory localization in this context, learning expected rewards of different signaling policies is a first step towards creating more intelligent, adaptive robot communication behaviors.

Considering the size of the data collection space, we assumed most positions could be physically reached by participants within three to four seconds. We also assumed participants would need approximately one second to notice the LED light cue and stop their sometimes noisy manipulation of LEGO pieces and find the robot. These assumptions were supported by the results of the experiment; the median and average (not normalized) response times were 4.0 seconds and 3.8 seconds, respectively. Given the above assumptions, the increasing and decreasing auditory signaling policies were held constant volume for one second initially and reached their constant end volumes after four seconds.

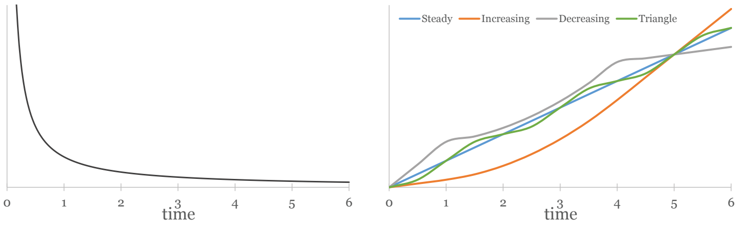
We calculated the average reward for each signaling policy and world state using the previously presented formula, as illustrated in Fig.8 (a). Surprisingly, we found that despite varying the minimum and maximum sound level intensities $L_{i\,min}$ and $L_{i\,max}$ to match the ambient sound levels, the learned policies π^* varied between our two ambient sound conditions (55 dbA vs 70 dbA) for interactions when the robot was out of the human's view. These results suggest that in the absence of initial visual cues, more salient or higher cost signals best support localization.

We found that the increasing sound policy performed the best across the most world states, despite having the longest average response time $t_{resp_{norm}}$. This result occurred due to a much lower cost term since this signaling policy always started at the minimal intensity, as shown in Fig.6 (3), which dominated the reward function. As participants typically took less than 5 seconds to reach the robot, this gave the increasing policy a significant advantange. The overall signaling scheme also caused the decreasing auditory policy to never be chosen as the ideal policy, despite yielding the second lowest average response time across states.

Due to the constant change in intensity throughout the triangle signal, we predicted that this policy would be the most salient. The triangle policy performed the best for the lowest visibility, highest ambient noise world states; however, we found that for many other world states, the triangle policy performed the worst. Since humans typically utilize intensity differences during localization [5], the constant variations in the triangle signal may have confused participants as several commented after the experiment that the sound was a bit "weird."

7 DISCUSSION

In this work, we investigated auditory signals for enabling humans to localize a co-located robot. First, we investigated the effect of broadband and tonal signals on people's ability to localize a robot during a collaborative task. Before this research, few works had



(a) Reward Term

(b) Cost Term

Robot Visibility (v)		Distance	Ideal Policy (π^*)	
Viewing Range	Obstructions	(d)	55 dbA	70 dbA
Out of view	Occluded	Close	Increasing	Triangle
Out of view	Occluded	Far	Steady	Triangle
In view	Occluded	Close	Increasing	Increasing
In view	Occluded	Far	Steady	Steady
Out of view	Not Occluded	Close	Increasing	Steady
Out of view	Not Occluded	Far	Increasing	Steady
In view	Not Occluded	Close	Increasing	Increasing
In view	Not Occluded	Far	Increasing	Increasing

(c)Ideal Policies

Figure 8: (a) The reward term based on response time; (b) The cost term based on response time and auditory signal, and (c) The auditory signal policies learned from the data collection for each world state.

explored auditory localization within the context of the everyday settings where robots are increasingly expected to operate.

Our initial user study allowed us to compare both sound variations to a baseline condition with no added auditory signal. We found that both types of added auditory signal significantly improved participants’ ability to accurately and quickly locate the robot. This improvement is logical as the robot’s naturally produced baseline sounds have low intensity levels compared to high ambient noise levels. Despite results from past works, we found that neither added auditory signal performed significantly better than the other, which may point to the need for greater research in the dynamic settings that robots are expected to operate in.

We also found that all three auditory conditions (baseline, tonal, and broadband) significantly affected participants’ ratings of the robot’s noticeability and annoyance. While participants rated the tonal and broadband conditions as significantly more localizable than the baseline, a post hoc analysis revealed that the two auditory signals were not rated significantly differently, matching the results shown by our objective metrics.

Finally, participants were asked to chose an auditory condition for each of these descriptors: most noticeable, most localizable, and most annoying. The tonal condition was chosen as both the most noticeable (88%) and most annoying (83%) auditory condition, while the broadband condition was chosen as the most localizable (67%). 79% of participants chose the broadband signal as their preferred auditory condition for future collaborations with a robot.

Overall, these results support the use of auditory signals for human localization of a co-located robot and the use of iconic broadband signals to minimize human annoyance. One of the weaknesses of this study was the removal of visual cues from the environment. Although this experimental setup enabled us to isolate the effects of the auditory signals, humans typically utilize both auditory and visual cues for localization in the real world. Moreover, the study also had mixed results in regards to humans’ ability to localize broadband sounds over pure tones.

To address these issues, we proposed an auditory signaling model using a Markov Decision Process (MDP) that incorporates information about the environment, including visual cues. The goal of this model is to enable a human to localize a robot while minimizing their annoyance with the robot’s auditory signals. The use of an MDP enabled the robot to account for uncertainty in the world, including the robot’s sensing capabilities and human’s physical perception of its signals.

As an initial application of this model, we performed a data collection with 18 participants. We learned ideal signaling policies for each world state and found that the increasing intensity auditory policy often yielded the greatest reward. Using our reward formulation, the increasing policy’s low sound intensity levels at earlier response times gave it an advantage over other explored policies given the typical participant response time.

Thus, a major limitation of this experiment is that the cost term dominated the reward function (1). Since the cost depends on the rate of change in intensity, altering this variable significantly affects the learned policies. The fast typical response time also caused the reward term to be minimized for many interactions. This shortcoming points to the need for better reward formulations that are more informed by human behavior.

In the data collection, we leveraged only four auditory signaling policies. As there is limited work in auditory localization, these results provide foundational insights and enables future exploration of a larger set of signaling policies, including other methods and rates of variation in signal intensity. In the future, we also plan to explore modulation of other properties of the signal (e.g., pitch) to create a more rich vocabulary of robot signals.

A primary strength of this work is that it is not tied to a specific sound; rather this work examines auditory icons holistically and examines how modulations of this class of sounds can improve localization of the robot. This generalized approach enables our results to be applied to a wide range of robot sounds with different designs and meanings. This work also adds to the growing body of literature that investigates iconic robot sounds while also employing these sounds in HRC tasks designed to mimic the real world [22, 34].

The proposed auditory signaling model also provides a generalized approach for intelligently planning robot communication behaviors while accounting for the inherent uncertainty in the world state. Although we currently apply this model to a simplified auditory localization problem that requires only a set of one-state MDPs, this formulation can be extended in the future to more complex problems that require more complex methods for solving, such as reinforcement learning.

These results offer a first step in creating intelligent auditory signaling policies that use iconic sounds to convey robot state information in an intuitive manner. Our findings further confirm that broadband sounds support localization with reduced annoyance, making them a useful communication tool for human-robot interaction. In the future, we plan to apply this model to a wider range of scenarios and use a human-subjects experiment to validate it against current robot policies that use static auditory signals.

ACKNOWLEDGMENTS

This work was supported by a NASA Space and Technology Research Fellowship (NSTRF) and NSF NRI IIS-1528121.

REFERENCES

[1] Rachid Alami, Alin Albu-Schäffer, Antonio Bicchi, Rainer Bischoff, Raja Chatila, Alessandro De Luca, Agostino De Santis, Georges Giralt, Jérémie Guiochet, Gerd Hirzinger, and others. 2006. Safe and Dependable Physical Human-Robot Interaction in Anthropic Domains: State of the Art and Challenges. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE.

[2] Kim Baraka, Ana Paiva, and Manuela Veloso. 2015. Expressive Lights for Revealing Mobile Service Robot State. In *Robot 2015: Second Iberian Robotics Conference*. Springer, 107–119.

[3] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of Nonverbal Communication on Efficiency and Robustness in Human-Robot Teamwork. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 708–713.

[4] Stephen A Brewster. 2002. *Non-Speech Auditory Output*. Lawrence Erlbaum Associates, 220–239.

[5] Douglas S Brungart and William M Rabinowitz. 1999. Auditory localization of nearby sources. Head-related transfer functions. *The Journal of the Acoustical Society of America* 106, 3 (1999), 1465–1479.

[6] Elizabeth Cha. 2018. Robot Sounds. http://lizcha.com/research/auditorylocalization/robotsounds. (2018).

[7] Elizabeth Cha and Maja Matarić. 2016. Using Nonverbal Signals to Request Help During Human-Robot Collaboration. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5070–5076.

[8] Bruce H Deatherage. 1972. Auditory and other sensory forms of information presentation. *Human Engineering Guide to Equipment Design* (1972), 123–160.

[9] Tilman Dingler, Jeffrey Lindsay, and Bruce N Walker. 2008. Learnabiltiy of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. In *International Conference on Auditory Display*.

[10] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. 2015. Effects of Robot Motion on Human-Robot Collaboration. In *ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 51–58.

[11] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and Predictability of Robot Motion. In *ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 301–308.

[12] Jill L Drury, Jean Scholtz, and Holly A Yanco. 2003. Awareness in Human-Robot Interactions. In *IEEE International Conference on Systems, Man and Cybernetics*, Vol. 1. IEEE, 912–918.

[13] Judy Edworthy and Neville Stanton. 1995. A user-centred approach to the design and evaluation of auditory warning signals: 1. Methodology. *Ergonomics* 38, 11 (1995), 2262–2280.

[14] Andy Farnell. 2010. *Designing Sound*. MIT Press, Cambridge MA.

[15] Harvey Fletcher and Wilden A Munson. 1933. Loudness, Its Definition, Measurement and Calculation . *Bell Labs Technical Journal* 12, 4 (1933), 377–430.

[16] Stavros Garzonis, Chris Bevan, and Eamonn O’Neill. 2008. Mobile Service Audio Notifications: intuitive semantics and noises. In *Australasian Conference on Computer-Human Interaction: Designing for Habitus and Habitat*. ACM, 156–163.

[17] William W Gaver. 1986. Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction* 2, 2 (1986), 167–177.

[18] Kazunori Kamewari, Masaharu Kato, Takayuki Kanda, Hiroshi Ishiguro, and Kazuo Hiraki. 2005. Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion. *Cognitive Development* 20, 2 (2005), 303–320.

[19] Peter Keller and Catherine Stevens. 2004. Meaning From Environmental Sounds: Types of Signal-Referent Relations and Their Effect on Recognizing Auditory Icons. *Journal of Experimental Psychology: Applied* 10, 1 (2004), 3–12.

[20] Günther Knoblich, Stephen Butterfill, and Natalie Sebanz. 2011. Psychological research on joint action: theory and data. *Psychology of Learning and Motivation-Advances in Research and Theory* 54 (2011), 59.

[21] John C Middlebrooks and David M Green. 1991. Sound localization by human listeners. *Annual Review of Psychology* 42, 1 (1991), 135–159.

[22] Dylan Moore, Hamish Tennent, Nikolas Martelaro, and Wendy Ju. 2017. Making Noise Intentional: A Study of Servo Sound Perception. In *ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 12–21.

[23] Bilge Mutlu, Allison Terrell, and Chien-Ming Huang. 2013. Coordination Mechanisms in Human-Robot Collaboration. In *ACM/IEEE International Conference on Human-Robot Interaction: Workshop on Collaborative Manipulation*.

[24] Konstantinos Papadopoulos, Kimon Papadimitriou, and Athanasios Koutsoklenis. 2012. The Role of Auditory Cues in the Spatial Knowledge of Blind Individuals. *International Journal of Special Education* 27, 2 (2012), 169–180.

[25] Roy D Patterson and TF Mayfield. 1990. Auditory warning sounds in the work environment. *Philosophical Transactions of the Royal Society of London: Biological Sciences* 327, 1241 (1990), 485–492.

[26] David R Perrott and Thomas N Buell. 1982. Judgments of sound volume: Effects of signal duration, level, and interaural characteristics on the perceived extensity of broadband noise. *The Journal of the Acoustical Society of America* 72, 5 (1982), 1413–1417.

[27] Peter Popoff-Asotoff, Jonathan Holgate, and John Macpherson. 2011. Which is Safer–Tonal or Broadband Reversing Alarms?). In *Acoustics*.

[28] Robin Read and Tony Belpaeme. 2012. How to Use Non-Linguistic Utterances to Convey Emotion in Child-Robot Interaction. In *ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 219–220.

[29] Robin Read and Tony Belpaeme. 2014. Non-Linguistic Utterances Should be Used Alongside Language, Rather than on their Own or as a Replacement. In *ACM/IEEE International Conference on Human-Robot Interaction Late Breaking Report*. ACM, 276–277.

[30] Julie Shah and Cynthia Breazeal. 2010. An Empirical Analysis of Team Coordination Behaviors and Action Planning With Application to Human-Robot Teaming. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52, 2 (2010), 234–245.

[31] Matthew Sneddon, Karl Pearsons, and Sanford Fidell. 2003. Laboratory study of the noticeability and annoyance of low signal-to-noise ratio sounds. *Noise Control Engineering Journal* 51, 5 (2003), 300–305.

[32] Daniel Szafir, Bilge Mutlu, and Terrence Fong. 2014. Communication of Intent in Assistive Free Flyers. In *ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 358–365.

[33] Daniel Szafir, Bilge Mutlu, and Terry Fong. 2015. Communicating Directionality in Flying Robots. In *ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 19–26.

[34] Hamish Tennent, Dylan Moore, Malte Jung, and Wendy Ju. 2017. Good Vibrations: How Consequential Sounds Affect Perception of Robotic Arms. In *IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 928–935.

[35] Luca Turchet, Simone Spagnol, Michele Geronazzo, and Federico Avanzini. 2016. Localization of self-generated synthetic footstep sounds on different walked-upon materials through headphones. *Virtual Reality* 20, 1 (2016), 1–16.