

Examining Audio Communication Mechanisms for Supervising Fleets of Agricultural Robots

Abstract—Agriculture is facing a labor crisis, leading to increased interest in fleets of small, under-canopy robots (agbots) that can perform precise, targeted actions (e.g., crop scouting, weeding, fertilization), while being supervised by human operators remotely. However, farmers are not necessarily experts in robotics technology and will not adopt technologies that add to their workload or do not provide an immediate payoff. In this work, we explore methods for communication between a remote human operator and multiple agbots and examine the impact of audio communication on the operator’s preferences and productivity. We develop a simulation platform where agbots are deployed across a field, randomly encounter failures, and call for help from the operator. As the agbots report errors, various audio communication mechanisms are tested to convey which robot failed and what type of failure occurs. The human is tasked with verbally diagnosing the failure while completing a secondary task. A user study was conducted to test three audio communication methods: earcons, single-phrase commands, and full sentence communication. Each user completed a survey to determine each method’s overall effectiveness and preferences. Our results suggest that the system using short phrases is the most positively perceived by participants and may allow for the human to complete the secondary task more efficiently.

Index Terms—Fleet Management, Robot Monitoring, Agricultural Robotics, Audio Communication

I. INTRODUCTION

Agriculture is currently facing a significant human labor crisis [1], harming profitability and causing negative downstream effects. As a result, precise actions (e.g., weeding, targeted pesticide spraying) are not feasible at the scale required for annual row crops (e.g., corn, soybeans), which dominate the Midwest and much of the US. In these settings, agriculture is only practical with heavy reliance on fertilizers, pesticides, and herbicides applied with large equipment (e.g., tractors, combines, planters) [2]–[4]. While this large equipment is familiar to farmers and can be automated to alleviate labor concerns [5], [6], such equipment is capital heavy, requires extra logistical oversight, introduces new safety risks to workers, and physically impacts the farm (e.g., soil-compaction, crop damage).

Small agricultural robots (agbots) can help alleviate the labor crisis and enable precision agriculture. These agbots are designed to be small, inexpensive, and intelligent, and have seen growing attention in recent years [7]–[9]. To fully address the labor shortage, these agbots must be both easy to use and able to be deployed at scale, where one human is supervising many robots.

Despite growing interest in agbots, farmers already tend to be overwhelmed with the large number of equipment and data sources that are constantly made available to them [10]. From

large suppliers automating their products to new technologies for data collection and interpretation [11], farmers are increasingly being driven to manage a large set of equipment with their own intricacies and interfaces [1]. Currently, the farmers take on the full brunt of orchestrating and monitoring the operation of their equipment, leading to a slow rate of adoption [10].

This application demonstrates the fact that developing novel technology does not guarantee that people will use or adopt the technology. Advances in robotics have a huge potential impact; however, if robots are not designed carefully, they will never successfully integrate into society or be accessible to non-experts [12]. We follow a Research through Design (RtD) methodology [13], [14], which suggests that a human-centered design process and an emphasis on *what* to design can lead to improved human-robot interaction tools [15]–[17]. Specifically, we intend to focus on identifying what interaction components should be developed to build an understanding of human-robot interaction and improve the rate of adoption [18]. Given rapid advances in Natural Language Processing and voice controlled robotics [19], we study what type of auditory interaction design most positively influences a user’s perception and productivity when using a remote robot monitoring system.

We consider situations where a fleet of agbots are deployed in a field and a remote operator supervises and monitors the robots. As the robots navigate up and down crop rows, there is some probability that they will encounter a failure that requires help from the human. When a failure is encountered, the robot control center will audibly prompt the human operator to provide assistance through verbal commands. Auditory interaction is well-studied as it has been shown to potentially boost efficiency and productivity of a user completing other visual tasks [20].

Handling failure scenarios is of critical importance for near-term deployment, as many agriculture tasks and environments are too complex for current agbots to handle without at least occasional failure [21]. Thus, when agbots are deployed humans will inevitably have to intervene. This setting follows the idea of sliding autonomy [22] and recent efforts to codify the levels of autonomy for field robots [23], which outlines how agbots with varying autonomous capabilities can interact with human operators [24]. In addition to practical motivations, failure cases are important design considerations as they have a strong effect on the user’s perceived competence and trustworthiness of the system [25].

This research provides insight into effectiveness and user

acceptance of audio communication interfaces when managing multiple autonomous robots. We studied the effect of earcons, single phrase commands, and full sentence speech on the user’s perception of the system and their efficacy in completing a secondary task. We present three contributions:

- 1) We develop a simulated control center to explore how humans interact and monitor agricultural robots deployed across a field, while potentially encountering failures that require human assistance.¹
- 2) We demonstrate how audio signals (either tones or natural language) can improve the remote operator’s efficiency and productivity compared to a traditional visual interface.
- 3) Our user study provides insight on how well an operator perceives various auditory interaction systems in a remote robot monitoring setting, indicating which system will most effectively be adopted.

This paper is organized as follows. We review relevant literature in Section II. In Section III, we present an overview of our exploratory study, our hypotheses, and our measures. Section IV presents the quantitative and qualitative findings from our user study. Finally, we discuss our conclusions in Section VI.

II. RELATED WORK

A. Audio in Design

Auditory interfaces can be defined as bidirectional, communicative connections between two systems [20]. There are various methods in which sound can be used for auditory interaction in human robot interaction (HRI). Sound can come as an utterance that it is deliberately used to communicate an intention including semantic-free utterances [26]. Sound can also come without intention such as consequential sound or movement sonification [15]. Using sound to intentionally convey information has many benefits including reducing visual overload, reinforcing visual messages, and providing additional information such as direction or emotion [20]. Our study uses the fact that sound can be used as a primary interface while the user is completing another visually intensive tasks, which best simulates a fleet supervision scenario. To the same degree that human-to-human interaction involves various senses such as sound, sight and touch, HRI will also involve various modes of interaction simultaneously. Multi-modal interaction is necessary for autonomous robots to seamlessly integrate into society [27], [28].

Many studies on audio communication study a user’s perception of robot speech, simple beeps, or consequential movement sound, ignoring the wide range of possibilities that auditory interaction can offer [29]. For the channel of communication from the robot to the human, there are four main ways in which data can be encoded into audio: auditory icons, earcons, sonification and speech [30]. Auditory icons are intuitive associations between a recognizable sound and a piece of information, earcons are unintuitive intentionally designed

associations, sonification maps information to variations in sound, and speech conveys information verbally [30]. Many further types of auditory interaction have been developed such as spearcons, sped up speech, [31] or audification, using data as sounds [20]. Earcons have often been studied in the HRI community in multimodal systems such as autonomous driving [32], adaptive automation of telerobotic control [28], healthcare [33], and many more applications [34]. In addition to the commonly used speech interface, we derive two other interfaces from the auditory interaction literature. One system is a mix of auditory icons and earcons which we simply refer to as earcons. The second is a shortened version of complete speech which we refer to as phrases. These systems should be able to convey the same amount of information to the user, yet are distinct enough to influence the user’s perception of the system and success with their secondary task.

B. Audio in Agriculture

A survey of 58 different publications with nontraditional human robot interactions in agriculture highlights some of the benefits of using speech [35]. In agricultural settings, audio interfaces tend to improved the usability of technology. Another recent HRI survey of over 50 papers determined that robots have not reached a level of design that allows for effective communication of faults by untrained users [12]. Literacy has been a major barrier preventing farmers who cannot read written instructions from using robots [35]. Many studies attempt to address this issue by surveying farmers and developing audio interfaces to assist under-educated farmers in communicating with robots and using technology [36]–[38]. These studies discovered that the complexity of a robot system is one of the main challenges farmers face in adapting technology, as employees lack the necessary skills to operate complicated robots. This conclusion underscores the importance of an intuitive and effective human-centered design when it comes to robot management on an autonomous farm.

Given the current state of autonomy in field and agricultural environments, it is unreasonable to expect fully autonomous systems that can handle all scenarios [35]. Therefore, research should focus on semi-autonomous and collaborative systems that work with human operators. Studying these systems must be done in a user-centered design fashion and is not only a technical engineering challenge, but must borrow insight from other domains such as psychology and sociology [35]. Our study builds on this insight to inform the design and guide the engineering development of robot monitoring systems.

C. Communicating Robot Failures Using Audio

Clear error communication is very important for a human and robot to overcome a misperception error and complete a task collectively [39]. Previous user studies on the failure rate in HRI tasks indicate that direct and clear communication is more important than dialogue [40]. This finding implies that task-oriented robots should focus more on concise speech or sounds than lengthy dialogue.

¹The codebase will be released upon acceptance.

However, completely restricting the robot’s vocabulary to simple commands can negatively impact the robot’s usability and the user’s perception. A previously conducted user study of a helper robot in the kitchen characterizes the concept of underperception and overperception [41]. In underperception, a human underestimates the capability of the robot, thus not utilizing it to its full extent. In overperception, the human overestimates the capability of the robot leading to the human’s expectations not being met, and therefore the human is less willing to work with the robot. All in all, they conclude that when a human misperceives a robot’s capabilities, they misuse it, and their acceptance of the robot decreases which can decrease the success of the collective HRI task [41].

Similar to the aforementioned design theories, the studies of underperception and overperception indicate that robotics research should focus on balancing technology with perception, since advanced robotic systems are pointless if the user does not perceive them correctly [41]. We extend this idea to the remote supervision and monitoring setting and focus on the user’s perception of auditory interaction. The results of [41] indicate that in a physical HRI setting speech improves perceived capability. However, in the remote operation setting, speech might improve perceived capability. Since the robot is not a humanoid machine directly interacting with the human, too much speech communication may be perceived as redundant and may not be vital to increasing the overall success in the human robot interaction. Thus, decreasing the speech capabilities at the control center may avoid overperception and increase the usability of the system.

A study investigating the effect of robot failure recovery strategies on the trust and perceived competence of a robot system, highlights why failure scenarios provide a unique opportunity to study human perception [25]. Failure scenarios are often when the human’s perception of the robot is affected the most. This motivates our farm simulation and why addressing robot failures may provide strong insight into the user’s perception of the system.

Most previous studies play videos of robots talking [25], [41] or have humans actually speak to a physical robot [40], [42]. We postulate that the remote robot monitoring setting can provide unique insight into how a human perceives the robot. For example, a user interacting with a humanoid robot may perceive a failure more negatively since they expect a human looking robot to act like a human. Whereas in a video, the user may perceive the robot less negatively since they do not see the robot in real life and are not physically interacting with it. The user’s perception of a remote robot fleet management system in failure scenarios is a mixture of both settings.

III. METHODS

A. Experimental Design

In order to study the design of an audio interface for agbot monitoring, we develop an autonomous farm environment using OpenAI gridworld simulation as shown in Figure 1. This simulation is meant to capture a fleet of robots navigating up

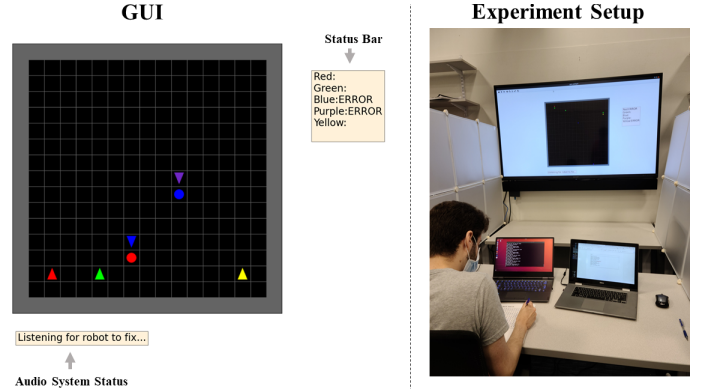


Fig. 1: *Left:* The GUI allows the user to visually deduce the state of each robot and the system. The status bar indicates which robots have failed, and those robots are shown on the grid stopped behind circles. In the instance shown above, the purple robot is stopped behind a blue circle (Unrecoverable Failure) and the blue robot is behind a red circle (Row Collision). The red, green and yellow robots are unobstructed and continuing to traverse their sections of the grid. The audio system status displays the state of the system. In this image, the system is waiting to hear a color indicating which robot to fix. *Right:* The experiment setup mimics the control center setting on an autonomous farm. The user is on an isolated desk in front of a TV screen with the GUI and is given the wordsearch puzzles to work on. They also have reference tables of the error solutions, earcon mappings, and scripts of each audio type (Tables III IV I).

and down crop rows. An image of the visual interface and mock control center is shown in Figure 1.

There are five robots represented as different colored triangles navigating up and down the columns of the grid sequentially, each traversing 1/5th of the grid and then halting to indicate their task is finished. There are 15 failure cases uniformly distributed on the grid represented as circles. When a failure is reached, the robot will stop and prompt the user to diagnose the error. Based on the error case, the human verbally communicates to the robot how to resolve the error and the robot will proceed. The failures, represented by circles, are not visible on the grid until one of the robots reach that failure, thus preventing the user from anticipating a failure.

We conduct four separate experiments on 13 users. The audio prompt that plays when a robot failure occurs is different in each of the four experiments: earcons, single phrases, full sentences, and no sound at all. A script of the interaction between the robots and the human is shown in Table I and provided to the user before each experiment. Before the participant begins the formal experiment, we completed a tutorial with them, where the researcher demonstrates how to fix the robots. The subject practices fixing the robots until they feel comfortable and confident in using the system. The purpose of this warm-up session is to minimize the effect that unfamiliarity has on the subject’s attitudes or quantitative

TABLE I: Scripts for each of the different audio communication modalities. When a failure is encountered, the robot communicates audibly and the human verbally responds with how to fix the failure.

Earcon	Phrase
Robot: “[red]” Human: “Fix the red robot.” Robot: “[robot_fail]” x 2 Human: “Navigate around.” Robot: “[robot_fixed]” Robot: “[blue], [green]” Human: “Fix the blue robot.” Robot: “[robot_fail]” x 1 Human: “Reverse and retry.” Robot: “[robot_fixed]” Robot: “Errors at green.” ...	Robot: “Error at red.” Human: “Fix the red robot.” Robot: “Untraversable obstacle.” Human: “Navigate around.” Robot: “Error fixed.” Robot: “Errors at blue, green.” Human: “Fix the blue robot.” Robot: “Row collision.” Human: “Reverse and retry.” Robot: “Error fixed.” Robot: “Errors at green.” ...
Sentence	
Robot: “There is an error at the red robot.” Human: “Fix the red robot.” Robot: “The red robot is facing an untraversable obstacle.” Human: “Navigate around.” Robot: “The error has been fixed.” Robot: “There are errors at the following robot blue, green.” Human: “Fix the blue robot.” Robot: “Row collision has occurred at the blue robot.” Human: “Reverse and retry.” Robot: “The failure has been fixed.” Robot: “There are still errors at the green robots.” ...	

success with the task. The audio system that was randomly chosen as the tutorial for a given subject was the system that they ran the experiment on last.

We measure the users perception of success and usability in the system through a survey after each experiment and their level of productivity through a secondary task score.

B. Error Cases




In an attempt to recreate realistic failures from the agricultural domain, we consider three common error cases: row collision, where the agbot collides with the crops; untraversable obstacle, where the agbot encounters an unexpected object in the field; and unrecoverable failure, where some fault or critical state has occurred. We assume that these failures can be reliably detected when the agbot is out in the field, using failure detection systems [43]. The row collision and untraversable obstacles are recoverable failures with proper guidance, whereas the unrecoverable failure requires a human operator to go to the field and rescue the robot. For the purposes of this study, all errors can be resolved from the operator’s commands. The errors are further described in Table II.

The user addresses each of the error cases using certain verbal commands. The user is given Table III as a reference during the experiment to know which commands to say for which error cases. Although the one-to-one mapping of errors

TABLE II: List of failure type with respective description and typical recovery solution.

Failure Type	
Row Collision	<i>Description:</i> The robot deviates from the center line and crashes into crops on either side of the row due to perception failures. <i>Solution:</i> Reverse and replan the path that tracks the center line.
Untraversable Obstacle	<i>Description:</i> The robot stops in front of the obstacle which obstructs the center line, but still has space to plan a collision free path. <i>Solution:</i> Navigate around the obstacle and then continue the robot’s original trajectory.
Unrecoverable Failure	<i>Description:</i> The path is fully blocked or the robot is in some failure scenario where it cannot continue without human intervention. <i>Solution:</i> Send a human to the field to help the robot recover.

TABLE III: User commands for the different failure modes.

Error #	GUI Icon	Error Type	Solution
1		Row Collision	“reverse and retry”
2		Untraversable Obstacle	“navigate around”
3		Unrecoverable Failure	“sending human”

to error fixes seems trivial and can easily be automated without a human operator, this system can be extended to a scenario where the human operator has more knowledge than the system and must make an informed decision about how to fix each failure. Even in its current state, the user chooses the order in which to fix multiple robots, which is not a trivial task for humans or planners [24]. Nonetheless, knowing how or which robot to address in our simulation is unrelated to the type of auditory prompt the user hears from the system, thus the triviality of the system does not discount the merit from a user-centered design perspective.

As this study focuses on the impact of audio communication on a fleet supervision task, the number and type of failure is kept constant so the failures themselves did not influence the user’s perception. There are exactly 15 errors in every simulation, 5 of every failure type. However, the location of each error is randomly sampled from a uniform distribution across the grid to simulate more realistic failure scenarios.

C. Audio Signals

The verbal interaction from the human to the robot remains constant throughout all the simulations, as described in Tables I and III. However, the auditory interaction from the robot to the human will change across each of the four experiments. There is the earcon system, the single-verbal phrase system, the full sentence speech, and the system with no sound. When one or more robots fail, it will prompt the user with audio indicating the colors of the failed robots. The user will have to verbally say a color of the robot they wish to fix. Next, the system will indicate which type of failure the robot is facing.

TABLE IV: Earcon Table. Each robot is associated with a unique sound that is related to its color. Another sound is played when the robot fails and when it is fixed.

Robot	Red	Green	Blue	Purple	Yellow
Sound	Siren	Leaves Rustle	Water Splash	Violin	Taxi Honk
Condition		Robot_fixed	Robot_fail		
Sound		Ready	Coin		

The user then must say the correct command to fix the system. Once fixed, the system will notify the user that the robot has been fixed and continue on with the simulation until another robot fails. Table I shows example scripts from each of the methods.

To keep the amount of information conveyed by each audio system constant, we developed a mapping from earcons to robot colors, shown in Table IV. To convey what type of failure the robot is at, the robot plays a coin noise to indicate the error number shown in Table III, once for row collision, twice for untraversable obstacle, and thrice for unrecoverable failure. Before the earcon experiment, an earcon tutorial program was ran which played each of the earcons and verbally said what they correspond to as described in Table IV.

For the no sound baseline, the users can only refer to the visual GUI to find out if a failure has occurred, which robot has encountered the failure, and what type of failure the robot is facing. The user had no audio prompt, thus they would have to occasionally look up from the secondary task to address the failures.

For all audio settings, the GUI shows a list of the failed robots. The failure type is shown within the grid world by the color of the circle at which the robot is stopped behind: red indicates row collision, green indicates untraversable obstacle, and blue indicates unrecoverable failure (see III). An example of the GUI (scaled down for clarity) is shown above in Figure 1.

The order in which the user went through the four different conditions was randomized and counterbalanced within subjects to reduce the bias. For example, if most people found the overall system difficult to use at the beginning and better by the end, randomizing the order should minimize the effect that plays on the results of the study.

D. Secondary Task and Productivity

In an autonomous farm control center setting, or any robot monitoring system, the user is likely not going to be fully focused on monitoring the robot. Instead, the operator would be completing other tasks and be prompted when a robot needs attention. Our experiment uses a wordsearch puzzle as a controlled secondary task. A wordsearch puzzle has a user find given words going in various linear directions in a grid of letters. The wordsearch puzzle is a good visual stimulus and cognitive load that the user can engage with as it does not

interfere with the auditory interaction of the system we wish to study. Many psychology studies use word search puzzles to study distraction or multitasking [44], [45].

Dividing the user’s attention across different senses is a simple and effective way to measure their productivity when switching attention between the two tasks at hand. The participant was put in a quiet constant environment, as shown in Figure 1, allowing us to isolate the grid simulation as the only audio stimuli the user was receiving. User studies with physical robots often have auditory background that could affect the user’s perception of the system [46]. However, our distraction-free environment isolates the audio’s effect on the user’s perception of the system and success in solving the wordsearch puzzles. Furthermore, the wordsearch puzzle is very easy for beginners to learn and does not give an advantage to those with more cultural knowledge or math practice (in contrast to crossword puzzles or math questions).

To score each audio condition, we counted how many words the participant found and divided by the total time it took them to run the simulation (the simulation stops when all 5 robots reach the end of their last crop row). This words per minute score is robust to small technical inconsistencies or pauses in the system as well as how the user decides to divide their attention in the system. If the system pauses or glitches, then the user has another second to think and find words. If the user only focuses on finding words, the number of failed robots will increase (and eventually come to a complete standstill), thus increasing the the time that the simulation takes to complete. On the other hand, if the user does not focus on finding words, they will receive a low productivity rate. Overall, the system encourages them to address both the wordsearch task as well as the robot failure task at the same time, which strengthens our findings on how audio affects efficiency when faced with a visually intensive task. At the beginning of the experiment, the users were made aware of this metric and the explanation of how it means they should focus on both the simulation and word search.

E. Participants

A total of 13 participants voluntarily performed the user study. Each of the simulations took approximately eight minutes to run, but with the explanations and tutorial the entire process took at most one hour. Every participant was a university affiliate; however, they had varying degrees of previous experience in robotics.

F. Survey Questions

The survey asked a few background questions before the experiment was conducted. After each experiment, the participant was asked to rate the following questions for each sound-simulation type on a 7 point Likert scale:

- Q1: *It was easy to diagnose and fix the errors in this system.*
Q2: *I was successful at guiding the robots passed their errors.*
Q3: *This system was overwhelming to use.*

The questions were created for the purpose of this design experiment. However, they relate to the competence dimension

of RoSAS [47] and the likeability dimension of Godspeed [48], which are two of the most common HRI metrics.

For each experiment, we asked the subject to report how many times they looked up in order to gauge how often they were using the GUI in each of the experiments. Between each round, the participant was allowed to change any of their previous answers. At the end of the survey, they were asked if they had changed any of their previous responses. This allowed the participant to reflect on past system experience, while giving us information about how their perceptions changed. After all conditions were tested, the participant was asked to rank the notification methods from 1 to 4 (best to worse) and to leave feedback and general impressions of the overall system.

G. Hypotheses

The following three hypotheses were developed and tested by the experiment setup.

H1: Any audio interface will facilitate better success, usability, and productivity over a purely visual method.

In a control center setting, the human will likely be addressing robot failures while operating other systems of the farm at once, which in our study is analogous to the word search puzzle. The sound notification thus acts as an interrupt, grabbing the user's attention only as necessary, which allows them to maximize the time they spend on the puzzle. With the purely visual interface, the user must look up at the GUI occasionally to check for errors, which may disrupt their focus more frequently and result in them negatively perceiving the system and performing worse on the secondary task.

During the control experiment with no sound prompts, the user gets to choose when to look up and address robot errors. One could argue that with sound notification a user's train of thought is interrupted, so looking up on their own might improve their performance. However, with sound notification, the user can fully address a robot's issue without ever looking up at the screen. Thus even if they lose their thought process, their eyes may still be on the word search and they still may be able to recognize words while addressing the failure.

H2: Single phrase communication provides the best user perceived success, usability, and capability of the system.

Speech capability is linked with improving the perception of social capability in a robot [41]. A human that can speak with a robot perceives the robot as more intelligent and competent, regardless of whether the robot is more physically able to complete the task. This has often been studied in a physical HRI task and can lead to more positive perception of a robot's abilities and thus risks overperception. However, when applied to a remote supervision setting, we hypothesize that speech no longer implies an improved perceived capability of a robot and thus will have no significant impact on the success of the task. Since the monitoring task has no observed physical embodiment, we are skeptical that the users will attribute social intelligence to speech.

H3: Single phrase communication will result in the most significant improvement in productivity when the user is completing a secondary task.

TABLE V: We report the averages for each measure. Productivity Score (Prod. Score) is in words per minute. Q1, Q2, and Q3 correspond to the average responses from the survey questions. Rank provides the average user ranking.

Audio Type	Prod. Score	Q1	Q2	Q3	Rank
Earcon	1.49	5.54	6.15	2.62	2.23
Phrase	1.73	6.54	6.77	1.85	1.38
Sentence	1.56	5.85	5.85	3.00	2.62
No Sound	1.44	3.85	5.46	4.62	3.77

TABLE VI: P-values of Paired T-tests. The p-values shown below indicate their is a significant difference ($\alpha = 0.05$) between the mean between the phrase system and both of the other audio systems for every question. Thus the phrase system has the most positive impact on the user's perception of the system and H2 is supported.

Question	Phrase and Earcon	Phrase and Sentence
Q1	0.003	0.041
Q2	0.013	0.008
Q3	0.013	0.014

We hypothesize that in the control center scenario the user will not want to be interrupted with a lengthy description of the problem nor hold a conversation with the system. However, they may appreciate some more informative and easily interpretable feedback than an earcon which they have to map to a robot color. Thus, the single-word communication system will likely provide the best balance between the two extremes for the human.

IV. RESULTS

The survey results are shown in Table V, columns Q1, Q2, and Q3. As expected, the condition with no sound performs the worst in every metric. Using the type of audio system as the treatment groups, ANOVA was performed on the results of each question, with the following findings: Q1 ($p < 0.001$), Q2 ($p = 0.03$), and Q3 ($p < 0.001$). Each results are shown to be significant with $\alpha = 0.05$ which supports H1.

To verify H2, paired T-tests were performed on each question of the survey results between the phrase system and the two other audio systems. A visual comparison of the results are shown in Figure 2. The survey results show that the user most positively perceives the system prompting them with single phrases. These results are statistically significant as shown by the p-values of the T-tests in Table VI. Looking at the rankings provided by the user (see Table V), phrase was ranked higher than each of the other methods, which indicates that the user had the most positive impression of the phrase system. H2 is supported by the data.

The results of the secondary task are aggregated in Table V, and the mean of the phrase system performs slightly higher than the rest. However, ANOVA was performed across the

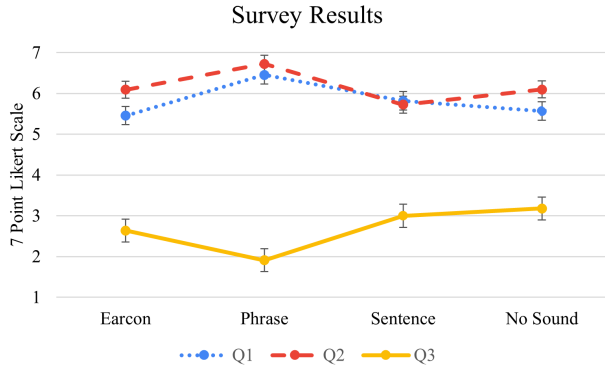


Fig. 2: Each line corresponds to a survey question and shows the mean score of each audio system on the 7-point Likert scale across the 13 participants. The phrase system is perceived as the easiest to use, most successful, and least overwhelming.

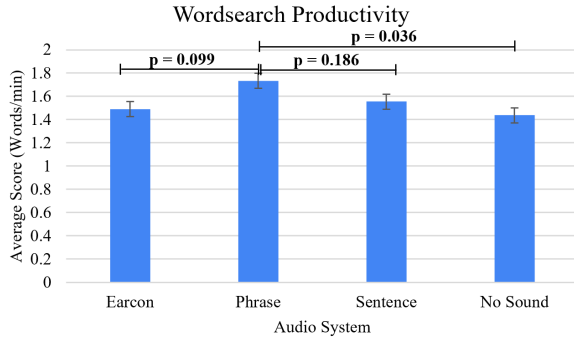


Fig. 3: The results show the means of the word search score and the p-values of paired T-tests between each system and the phrase system. This figure indicates a trend towards a the single phrase audio system being the most productive, however, statistical significance is only shown in comparison with No Sound.

different auditory interaction systems and indicated that the results are not statistically significant ($p = 0.78$). Nonetheless, we perform tailed T-tests with the phrase system and every other system and display the significance values in Figure 3.

V. DISCUSSION

The results of this study can help inform the design of a fleet management system for agricultural robots. The result of H1 answers the motivating question of the study concerning whether or not audio is necessary and beneficial in remote robot management. We have proven that in our use-case, an audio interface can make a significant difference in the way in which the user perceives the system. Although agbots with high levels of autonomy are rapidly being deployed, the insights from this study can help inform the design of robot monitoring systems in other applications. Remote operation is also being used in the healthcare [49] and manufacturing [50]. Understanding the methods of how the system can best inform the user would allow these applications to become easier to use

for the general population and would allow robots to become more widespread. More humans could be using more robots at the same time, while still being able to focus on other tasks.

After testing three separate audio systems, we can further conclude that the phrase system positively affected peoples' perception of the system through the data's support of H2. They found it to be easiest to use, most successful, and least overwhelming. Furthermore, they rated it the highest which means it's the best suited system for completing a task HRI task with a robot monitoring system from a user-centered standpoint. One explanation of these results is that the system does not need full natural language processing ability because if it did, the user might overperceive the system as capable of more than it actually is. These unrealistic expectations would increase the chances of a failure of the HRI task as the human will likely try to have the system do more than it is capable of, similar to what was discovered in [41]. Imagine the scenario where a human operator with minimal robotics experience is operating robots on a farm remotely. If the robots are speaking full sentences back to the human, the human will think the system is capable of more than it is truly capable of. Thus, the human will try to speak full sentences back. It may try to test and interact with the system in ways that the developer of the system did not intend. In this scenario, the system will probably fail to understand the human correctly and either perform the task incorrectly or not perform the task at all. Instead, if the system only speaks word or phrase level commands, the user perceives the system as less capable, thus the user's perception of the systems capability better matches it's true capability.

Another explanation of why the phrase system scored the best in the surveys is that people got frustrated with the other systems. Theoretically, each system required a slightly larger cognitive load on the user in comparison to the phrase system. The earcon system required the user to learn the mapping between the earcons and the color of the robots. Most participants were learning this earcon to color association during the experiment. Instead of looking at the GUI to see the color of the robots, the user would look at the earcon table (Table IV) and determine which robot to fix based on the earcon. Eventually, they were able to learn the mappings and did not have to look away from the wordsearch to fix robot errors. This difference in cognitive load might have influenced the results of the study. There might be a more intuitive mapping than earcons to color. For example, the user may have found it easier to indicate which robot to fix by saying the name of the sound instead of the color it matches. Nonetheless, users preferred the earcon system over the full sentence system, and generally found the earcon system to be more successful and less frustrating. One of the user's described that "it was helpful to have an association between the sounds and the robots that needed fixing." However, all verbal comments about the sentence system given by the participants were negative. Furthermore, if we remove the worst performing subjects (the outliers) from our analysis, the earcon system in fact performs better than the sentence system for

productivity and a few of the survey questions. This indicates that earcon and sentence conditions were very similar from the user's perspective, however, the phrase methodology was much more positively regarded than both the other methods.

Users often seemed frustrated with the sentence system because it required a larger cognitive load. One user reported that they would "never want a system that speaks in full sentences as it is highly frustrating". The problem with the sentence method was that in order to communicate one small piece of information, the system takes many more seconds than necessary. Such system behavior distracted the user from their task and often led the user to speak before the system was done speaking. When the system did not register their response, they had to speak again and became frustrated. Overall, comprehending the full sentence and having to be careful of when they can speak increased the cognitive load on the user.

Finally, the fully visual system also involved an increased cognitive load compared to the phrase system as the user always had to take their eyes off the wordsearch to address the issue. They also had to be aware enough to remember to occasionally look up, which means their mind can never fully focus on the wordsearch.

The increase in cognitive load from the no sound, earcon, and full sentence conditions is most likely the reason why the single phrase mechanism was the best. This idea should not be considered in isolation. When creating robot monitoring designs that the user may be multitasking with, communicating the most information in the least amount of sound is evidently a good design choice. However, too little sound like the earcon system requires the user to still do some work themselves so the system has to take that into account.

To our surprise, our devised secondary task and its related metric did not produce significant results, leaving H3 unsupported. We hypothesized that the phrase system will improve the user's capability and increase productivity. Although, H3 was not statistically supported, we do see that on average the phrase system provides a higher productivity (see Figure 3). These results can inform the design of such a system and future studies that wish to investigate this further. One simple explanation for the lack of significance is the sample size that was tested, which can be expanded in future works.

Another explanation is that the imperfections in the speech recognition system had a large negative effect on the productivity of each system, making the potential increase in productivity that could have occurred across the systems insignificant. One drawback of our design, was that the robot monitoring system occasionally misheard what the human said. For example, it would sometimes hear "riversand retry" instead of "reverse and retry" or "fix the rad robot" instead of "fix the red robot." Such mistakes required the user to say the command repeatedly until the robot understood them. This problem was often exacerbated by participants who's first language was not English. Regardless of which condition the user was experiencing, if the robot frequently mishears them and the user must repeat themselves, productivity will

be decreased across all systems significantly, thus negating the increase in productivity from the different audio prompts.

VI. CONCLUSION

In this paper, we studied how remote supervision of an agricultural robot through speech and audio signals affects a user's perception of the system and productivity in a secondary task. Understanding this relationship would allow robotics research to focus on developing systems that match a user's perception to improve the overall human-robot interaction. The results indicate that the average user is more likely to find the single word command interface the easiest to use. This provides the best balance between perception and capability in the case of fleet management for an autonomous farm.

Although productivity was not improved significantly by the phrase system, it did improve user perception. This result indicates that user perception in this remote robot management scenario had a more noticeable effect on informing the design of the system than the user's productivity with the system, which underscores the necessity of human-centered design. We found that users preferred a system that communicated with simple phrases, followed by simple sound communication. Full sentence communication was ranked last by participants. These findings match our intuition that full natural language understanding capabilities may not be needed in remote robot monitoring systems as people often found them more annoying than helpful.

Understanding the optimal interface and communication mechanisms for agricultural robots will help design technology and robots that are easy to use by non-experts. What seems intuitive to an academic, may not be as easy to use for a person with minimal robotics experience. Thus, the applied survey methodology is perfect for alleviating these differences in design and assuring that the technology that is developed is accessible to all types of users. By improving the design of such fleet management systems, we may increase the likelihood of adoption, paving the way for agbots to be deployed at scale.

REFERENCES

- [1] California Farm Bureau Federation, "Survey: California farms face continuing employee shortages," <https://www.cfbf.com/news/survey-california-farms-face-continuing-employee-shortages/>, 2019, accessed: Feb. 17, 2020.
- [2] A. J. Capellesso, A. A. Cazella, A. L. Schmitt Filho, J. Farley, and D. A. Martins, "Economic and environmental impacts of production intensification in agriculture: comparing transgenic, conventional, and agroecological maize crops," *Agroecology and Sustainable Food Systems*, vol. 40, no. 3, pp. 215–236, 2016.
- [3] J. A. Foley, N. Ramankutty, K. A. Brauman, E. S. Cassidy, J. S. Gerber, M. Johnston, N. D. Mueller, C. O'Connell, D. K. Ray, P. C. West, C. Balzer, E. M. Bennett, S. R. Carpenter, J. Hill, C. Monfreda, S. Polasky, J. Rockstram, J. Sheehan, S. Siebert, D. Tilman, and D. P. M. Zaks, "Solutions for a cultivated planet," *Nature*, vol. 478, no. 7369, pp. 337–342, 2011.
- [4] H. C. J. Godfray, J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, J. F. Muir, J. Pretty, S. Robinson, S. M. Thomas, and C. Toulmin, "Food security: The challenge of feeding 9 billion people," *Science*, vol. 327, no. 5967, pp. 812–818, 2010.
- [5] B. V. Ortiz, K. Balkcom, L. Duzy, E. Van Santen, and D. Hartzog, "Evaluation of agronomic and economic benefits of using rtk-gps-based auto-steer guidance systems for peanut digging operations," *Precision agriculture*, vol. 14, no. 4, pp. 357–375, 2013.
- [6] B. Erickson and D. Widmar, "2015 precision agricultural services dealership survey results," *Staff Paper*, no. 3-10, 2015.
- [7] S. Yaghoubi, N. A. Akbarzadeh, S. S. Bazargani, S. S. Bazargani, M. Bamizan, and M. I. Asl, "Autonomous robots for agricultural tasks and farm assignment and future trends in agro robots," *International Journal of Mechanical and Mechatronics Engineering*, vol. 13, no. 3, pp. 1–6, 2013.
- [8] S. M. Pedersen, S. Fountas, H. Have, and B. Blackmore, "Agricultural robots: system analysis and economic feasibility," *Precision agriculture*, vol. 7, no. 4, pp. 295–308, 2006.
- [9] J. Billingsley, A. Visala, and M. Dunn, "Robotics in agriculture and forestry," in *Springer handbook of robotics*. Springer, 2008, pp. 1065–1077.
- [10] AgWeb Farm Journal, "Avoid equipment and technology downtime this spring," <https://www.agweb.com/article/avoid-equipment-and-technology-downtime-this-spring-NAA-ben-potter>, 2016, accessed: Feb. 20, 2020.
- [11] D. Vasisht, Z. Kapetanovic, J. Won, X. Jin, R. Chandra, S. Sinha, A. Kapoor, M. Sudarshan, and S. Stratman, "Farmbeats: An IoT platform for data-driven agriculture," in *14th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 17)*, 2017, pp. 515–529.
- [12] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in psychology*, vol. 9, p. 861, 2018.
- [13] J. Zimmerman, J. Forlizzi, and S. Evenson, "Research through design as a method for interaction design research in hci," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007, pp. 493–502.
- [14] M. Luria, J. Zimmerman, and J. Forlizzi, "Championing research through design in hri," *arXiv preprint arXiv:1908.07572*, 2019.
- [15] F. A. Robinson, M. Velonaki, and O. Bown, "Smooth operator: Tuning robot perception through artificial movement sound," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2021, pp. 53–62.
- [16] M. Hoggemüller, W.-Y. Lee, L. Hespanhol, M. Tomitsch, and M. Jung, "Beyond the robotic artefact: Capturing designerly hri knowledge through annotated portfolios," in *1st international workshop on Design-erly HRI Knowledge. Held in conjunction with the 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2020.
- [17] M. Luria, S. Reig, X. Z. Tan, A. Steinfeld, J. Forlizzi, and J. Zimmerman, "Re-embodiment and co-embodiment: Exploration of social presence for robots and conversational agents," in *Proceedings of the 2019 on Designing Interactive Systems Conference*, 2019, pp. 633–644.
- [18] M. L. Lupetti, C. Zaga, and N. Cila, "Designerly ways of knowing in hri: broadening the scope of design-oriented hri through the concept of intermediate-level knowledge," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2021, pp. 389–398.
- [19] P. Chang, S. Liu, H. Chen, and K. Driggs-Campbell, "Robot sound interpretation: Combining sight and sound in learning-based control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.
- [20] S. C. Peres, V. Best, D. Brock, C. Frauenberger, T. Hermann, J. G. Neuhoff, L. Nickerson, B. Shinn-Cunningham, and A. Stockman, "Auditory interfaces," *HCI beyond the GUI: design for haptic, speech, olfactory, and other nontraditional interfaces*, pp. 147–195, 2008.
- [21] J. P. Vasconez, G. A. Kantor, and F. A. A. Cheein, "Human-robot interaction in agriculture: A survey and current challenges," *Biosystems engineering*, vol. 179, pp. 35–48, 2019.
- [22] M. B. Dias, B. Kannan, B. Browning, E. Jones, B. Argall, M. F. Dias, M. Zinck, M. M. Veloso, and A. Stentz, "Sliding autonomy for peer-to-peer human-robot teams," in *Proceedings of the international conference on intelligent autonomous systems*, 2008, pp. 332–341.
- [23] G. Chowdhary, C. Soman, and K. Driggs-Campbell, "Levels of autonomy for field robots," <https://www.earthsense.co/news/2020/7/24/levels-of-autonomy-for-field-robots>, 2020.
- [24] G. Swamy, S. Reddy, S. Levine, and A. D. Dragan, "Scaled autonomy: Enabling human operators to control robot fleets," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 5942–5948.
- [25] S. Reig, E. J. Carter, T. Fong, J. Forlizzi, and A. Steinfeld, "Flailing, hailing, prevailing: Perceptions of multi-robot failure recovery strategies," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2021, pp. 158–167.
- [26] S. Yilmazyildiz, R. Read, T. Belpeame, and W. Verhelst, "Review of semantic-free utterances in social human-robot interaction," *International Journal of Human-Computer Interaction*, vol. 32, no. 1, pp. 63–85, 2016.
- [27] J. J. Steil, F. Röthling, R. Haschke, and H. Ritter, "Learning issues in a multi-modal robot-instruction scenario," in *Workshop on Imitation Learning, Proc. IROS*. Citeseer, 2003.
- [28] D. B. Kaber, M. C. Wright, and M. A. Sheik-Nainar, "Investigation of multi-modal interface features for adaptive automation of a human-robot system," *International journal of human-computer studies*, vol. 64, no. 6, pp. 527–540, 2006.
- [29] C. Frauenberger, T. Stockman, and M.-L. Bourguet, "A survey on common practice in designing audio in the user interface," in *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21*, 2007, pp. 1–9.
- [30] D. K. McGookin and S. A. Brewster, "Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition," *ACM Transactions on Applied Perception (TAP)*, vol. 1, no. 2, pp. 130–155, 2004.
- [31] B. N. Walker, A. Nance, and J. Lindsay, "Spearcons: Speech-based earcons improve navigation performance in auditory menus." Georgia Institute of Technology, 2006.
- [32] N. Gang, S. Sibi, R. Michon, B. Mok, C. Chafe, and W. Ju, "Don't be alarmed: Sonifying autonomous vehicle perception to increase situation awareness," in *Proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications*, 2018, pp. 237–246.
- [33] G. Rosati, A. Rodà, F. Avanzini, and S. Masiero, "On the role of auditory feedback in robot-assisted movement training after stroke: review of the literature," *Computational intelligence and neuroscience*, vol. 2013, 2013.
- [34] G. Johannsen, "Auditory displays in human-machine interfaces," *Proceedings of the IEEE*, vol. 92, no. 4, pp. 742–758, 2004.
- [35] A. Rodríguez, A. Fernández, and J. H. Hormazábal, "Beyond the gui in agriculture: a bibliographic review, challenges and opportunities," in *Proceedings of the International Conference on Human Computer Interaction (HCI)*, 2018, pp. 1–8.
- [36] S. Ghosh, A. Garg, S. Sarcar, P. S. Sridhar, O. Maleyvar, and R. Kapoor, "Krishi-bharati: an interface for indian farmer," in *Proceedings of the IEEE Students' Technology Symposium*, 2014, pp. 259–263.
- [37] K. Bali, S. Sitaram, S. Cuendet, and I. Medhi, "A hindi speech recognizer for an agricultural video search application," in *Proceedings of the 3rd ACM Symposium on Computing for Development*, 2013, pp. 1–8.
- [38] F. Redhead, S. Snow, D. Vyas, O. Bawden, R. Russell, T. Perez, and M. Brereton, "Bringing the farmer perspective to agricultural robots," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 2015, pp. 1067–1072.

- [39] N. Schütte, B. Mac Namee, and J. Kelleher, "Robot perception errors and human resolution strategies in situated human-robot dialogue," *Advanced Robotics*, vol. 31, no. 5, pp. 243–257, 2017.
- [40] K. Fischer, B. Soto, C. Pantofaru, and L. Takayama, "Initiating interactions in order to get help: Effects of social framing on people's responses to robots' requests for assistance," in *the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 2014, pp. 999–1005.
- [41] E. Cha, A. D. Dragan, and S. S. Srinivasa, "Perceived robot capability," in *the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2015, pp. 541–548.
- [42] H. Hüttenrauch and K. Severinson-Eklundh, "To help or not to help a service robot: Bystander intervention as a resource in human-robot collaboration," *Interaction Studies*, vol. 7, no. 3, pp. 455–477, 2006.
- [43] T. Ji, S. T. Vuppala, G. Chowdhary, and K. Driggs-Campbell, "Multi-modal anomaly detection for unstructured and uncertain environments," *arXiv preprint arXiv:2012.08637*, 2020.
- [44] N. Harrell, J. McKulka, C. Kremm, and B. Mitchell, "Analysis of gaze on word search puzzles."
- [45] K. McMahon, B. Sparrow, L. Chatman, and T. Riddle, "Driven to distraction: The impact of distracter type on unconscious decision making," *Social Cognition*, vol. 29, no. 6, pp. 683–698, 2011.
- [46] E. Martinson and D. Brock, "Improving human-robot interaction through adaptation to the auditory scene," in *Proceedings of the ACM/IEEE international conference on Human-robot interaction (HRI)*, 2007, pp. 113–120.
- [47] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas) development and validation," in *Proceedings of the ACM/IEEE International Conference on human-robot interaction (HRI)*, 2017, pp. 254–262.
- [48] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International journal of social robotics*, vol. 1, no. 1, pp. 71–81, 2009.
- [49] G. Yang, H. Lv, Z. Zhang, L. Yang, J. Deng, S. You, J. Du, and H. Yang, "Keep healthcare workers safe: application of teleoperated robot in isolation ward for covid-19 prevention and control," *Chinese Journal of Mechanical Engineering*, vol. 33, no. 1, pp. 1–4, 2020.
- [50] L. Wang, "Collaborative robot monitoring and control for enhanced sustainability," *The International Journal of Advanced Manufacturing Technology*, vol. 81, no. 9, pp. 1433–1445, 2015.