**Motivation:** Human-computer interaction is dominated by visual and tangible interfaces, i.e. touch screens or buttons[1]. Common audio interfaces are computationally heavy intelligent assistants, such as Siri, Google, and Alexa, which do not translate well to task-specific applications[1] where dialogue can be minimized, and usability can remain. Robotics research is rapidly adapting audio communication interfaces, however, a divide in the literature highlights this computation-usability oversight. Papers either focus on enhancing the robot's speech accuracy or analyzing the human's perception of the robot. Very little research looks at a specific context and addresses the interplay between how accurately the robot perceives the human's speech input and how well the human perceives the robot's audio output. I intend to study this dynamic to maximize the effectiveness of any task a human and robot collectively partake in using a speech interface.

I plan to research this phenomenon in agriculture as a starting point for task-oriented embedded robotic applications. Food scarcity has become a dominant concern as many farmworkers have left the field and food production may not meet the rising demands of urban areas[2,3]. Studies conducted in Human-Robot Interaction (HRI) conclude that some tasks in agriculture are too complex for robots, such as pruning and thinning, so HRI should focus on the efficiency and profitability of robots that assist humans in these tasks[2].

Imagine the scenario where a robot is used to aid a group of laborers on a farm. The robot follows them around the field as the laborers pick crops and load them onto the robot. Once full, it autonomously navigates back to the farmhouse, delivers the load, and returns to the field for the next load. Examining this scenario, we can see the benefits of an audio interface. The robots need to be operated, but the workers are preoccupied with the crops. If the robot has an issue that needs to be communicated, introducing a touch screen or monitor would be excessive, impractical, and costly in a farm setting. In addition, this robot may include lidars, depth sensors, cameras and computational power for collision avoidance, crop counting, and numerous other functionalities; thus, we must be wary of the audio interface's computational resources.

**Intellectual Merit:** The issue is that most of the literature in Automatic Speech Recognition (ASR) describes high-performance applications using Large Vocabulary Conversational Speech Recognition (LVCSR), which maps audio signals to about 50,000-100,000 vocabulary words[4]. This would provide unnecessary computational overhead for many embedded robotic applications. In the agriculture example, many words would never be used based on the simple commands needed for the robot start, stop, and communicate its destination or errors. Previous user studies show that direct and clear communication is more important than dialogue when it comes to studying the failure rate in HRI tasks[5]. This implies that task-oriented robots should focus more on concise speech than lengthy dialogue. However, if you completely restrict the robot's vocabulary to simple commands, this can negatively impact the robot's usability and the user's perception. 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[6].

My goal is to research and create various speech communication models between a human and a robot and characterize the most effective and natural interface. I hypothesize that the performance and usability of a task-oriented robot will improve through an audio interface by limiting its ASR resource costs while maintaining sufficient levels of user perception. There is much literature on low resource ASR using LVCSR for handheld devices or humanoid robots, but little work has been done addressing low resource ASR for task-oriented robots while considering its effect on perception, especially in agriculture.

**Methods:** My research will follow 4 steps:

**1.** *Create a metric to gauge human perception and computational efficacy:* Human perception will be observed physiological and behavioral indicators as well as self-reported metrics through surveys. Computational efficacy will be measured in the software by analyzing the algorithms space and time complexity compared to its accuracy. For hardware-based metrics, I will compute the ASR model's computational and representational costs in terms of bit precision and multiply-accumulates.

**2.** *Conduct a user study testing human perception of a given ASR model:* I will begin by testing the canonical ASR Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) using the Kaldi ASR toolkit[7] on the EarthSense TerraSentia robot. During these experiments, I will measure the user's perception of the robots thrugh the determined metrics in (**1**). In 2017, I conducted a user perception study with a statistically significant result ($p = 0.0208$) showing that haptic feedback instead of visualization of an object increases

the agency a user feels over their hands in virtual reality. After that experience, I am confident I will be able to implement a method to test the human perception of speech interfaces in HRI.

**3.** *Analyze the ASR model's computational efficacy and improve on the model:* I will try to decrease the computational overhead given by the metric determined in (**1**), through altering the model. For example, the HMM-GMM models trained in (**2**) would use LVCSR. As per my hypothesis, this provides a vocabulary that is larger than necessary for the agriculture robot assistant task, so I will switch the model to use a small vocabulary approach[1] and test to see if the model upholds the same accuracy.

**4.** *Iterate*: I will continue to iterate between (**2**) and (**3**), testing the human's perception of the robot and then testing and improving the robot's computational efficiency until I find a model with a high human perception and ASR accuracy but low computational overhead. There are numerous models or alterations I could test. For example, I could implement a low-resource ASR method such as a Deep Maxout Network[8] instead of HMM-GMM. If a learned model is excessive for a small vocabulary, or the perception studies in (**2**) indicate that the speech can be limited to only a handful of phrases, I could create an algorithm to map the Mel-Frequency Cepstral Coefficients of speech phonemes to robot actions.

**Broader Impacts**: When I researched agriculture robotics in 2018, I noticed some oversight that theoretical research had on practical applications. I worked on autonomous navigation, while others in my lab worked on computer vision-based apple counting and reinforcement learning based strawberry picking. For these tasks, we operate the robots in a lab using laptops or remote controllers, which is not feasible for farmers in the field. This gives me the motivation to research an audio interface between the human and robot to the point of commercialization, allowing robots to become more accessible and commonplace in society.

A 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[3]. In agriculture, literacy has been a major barrier preventing farmers who cannot read written instructions from using robots[9]. My research intends to create an intuitive yet effective audio interface, so the use of robots is not constrained to a Ph.D. researcher with technical expertise and expensive lab equipment.

Lowering the communication barrier between humans and robots would allow for widespread production and adoption of robots which would give the U.S a competitive economic edge. In agriculture, improving yield would help fight global food scarcity. In factories, we would get a better output increasing Gross Domestic Product. Small businesses that cannot afford expensive robotic equipment and expertise will be able to use a task-oriented robot to carry a load, package boxes, etc. In everyday interactions, robotic systems could become more commonplace with a simple yet effective audio interface that understands what it needs to as opposed to an entire human language. Society will be able to produce more goods and services as a team alongside robots, instead of being replaced by them.

**References**: [1] Doukas, N., et al. Current trends in small vocabulary speech recognition for equipment control. AIP Conference Proceedings 1872, 020029. (2017) [2] Vasconez, J. P., et al. Human–robot interaction in agriculture: A survey and current challenges. Biosystems Engineering, 179, 35-48. (2019) [3] Honig, S., et al. Understanding and resolving failures in human-robot interaction: Literature review and model development. Frontiers in psychology, 9, 861. (2018). [4] Meteer, Marie. "Choosing the Right Technology for your Speech Analytics Project," http://callminer.com/wp-content/whitepapers/The-Right-Technology-for-Speech-Analytics.pdf [5] Fischer, K., et al. Initiating interactions in order to get help: Effects of social framing on people's responses to robots' requests for assistance. The 23rd IEEE International Symposium on Robot and Human Interactive Communication. 999-1005 (2014). [6] E. Cha, et al. Perceived robot capability. 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). 541-548 (2015). [7] Povey, D., et al. "The Kaldi speech recognition toolkit." http://www.danielpovey.com/files/2011_asru_kaldi.pdf. [8] Miao, Y., et al. Deep maxout networks for low-resource speech recognition. IEEE Workshop on Automatic Speech Recognition and Understanding. 398-403 (2013). [9] Rodríguez, A., et al. Beyond the GUI in agriculture: a bibliographic review, challenges and opportunities. Proceedings of the XIX International Conference on Human-Computer Interaction. 1-8 (2018).