Prosody as a Teaching Signal for Agent Learning: Exploratory Studies and Algorithmic Implications

Matilda Knierim*
Vrije Universiteit
Amsterdam, Netherlands

Sahil Jain*
Sony AI
Frisco, Texas, USA

Murat Han Aydoğan
Koç University
Istanbul, Turkey

Kenneth Mitra
University of Texas at Austin
Austin, Texas, USA

Kush Desai
University of Texas at Austin
Austin, Texas, USA

Akanksha Saran†
Sony AI
San Francisco, California, USA
akanksha.saran@sony.com

Kim Baraka†
Vrije Universiteit
Amsterdam, Netherlands
k.baraka@vu.nl

ABSTRACT

Agent learning from human interaction often relies on explicit signals, but implicit social cues, such as prosody in speech, could provide valuable information for more effective learning. This paper advocates for the integration of prosody as a teaching signal to enhance agent learning from human teachers. Through two exploratory studies—one examining voice feedback in an interactive reinforcement learning setup and the other analyzing restricted audio from human demonstrations in three Atari games—we demonstrate that prosody carries significant information about task dynamics. Our findings suggest that prosodic features, when coupled with explicit feedback, can enhance reinforcement learning outcomes. Moreover, we propose guidelines for prosody-sensitive algorithm design and discuss insights into teaching behavior. Our work underscores the potential of leveraging prosody as an implicit signal for more efficient agent learning, thus advancing human-agent interaction paradigms.

CCS CONCEPTS

• **Computing methodologies** → **Learning from implicit feedback**; *Learning from demonstrations*; • **Human-centered computing** → Empirical studies in HCI; **Auditory feedback**; *User studies.*

KEYWORDS

Human-robot/agent interaction; Machine learning; Social signals

*Both authors contributed equally to this research.
†Equal advising.

ICMI '24, November 4–8, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685735

ACM Reference Format:
Matilda Knierim, Sahil Jain, Murat Han Aydoğan, Kenneth Mitra, Kush Desai, Akanksha Saran, and Kim Baraka. 2024. Prosody as a Teaching Signal for Agent Learning: Exploratory Studies and Algorithmic Implications. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3678957.3685735

1 INTRODUCTION

Recent years have witnessed a surge in research focused on how agents can learn from human interactions, predominantly concentrating on clear and overt cues such as natural language feedback [14, 25, 26]. Nonetheless, human communication is inherently complex, infused with a variety of subtle and implicit signals that, we hypothesize, could significantly enhance the agent's learning process. The field of interactive learning with multi-modal human cues [26, 28] has started leveraging implicit signals such as clicker-based feedback (perfect and imperfect) [11, 23, 55], eye movements [38–40, 53], facial expressions [9, 27], gestures [27, 50], haptic feedback [5, 6], and object and environmental sounds [3, 10, 12, 52].

In this emerging research landscape, one modality stands out as clearly underexplored. *Prosody*, which involves various acoustic properties of speech such as tone, pitch, rhythm, and intonation, plays a critical role in human-human and human-animal interactions, where it serves as a key vehicle for conveying emotions, intentions, and expectations [13, 44]. For example, a rising intonation at the end of a sentence can distinguish an assertion from a question, or varying stress on words can alter the meaning of a phrase, demonstrating the nuanced role of prosody in communication. Despite its clear importance, prosody has not been extensively studied as a teaching signal within the realm of agent learning. This paper seeks to bridge this gap by delving into the potential of prosody to act as an informative signal for agents to learn from. By examining how prosody can aid in the interpretation and understanding of verbal instructions, we aim to underscore its value not only in facilitating agent learning but also in enriching our general understanding of human interaction, thereby contributing to more nuanced and effective communication models.

This paper considers the setting where a human who is an expert at the task uses speech as a teaching modality to teach an agent a near-optimal policy. We specifically consider two scenarios: one where a reinforcement learning (RL) agent needs to learn a policy from binary speech feedback (Sec. 3), and the other where an imitation learning agent learns from speech-augmented demonstrations, i.e., the expert talks while providing full demonstrated trajectories (Sec. 4). For both studies, we focused only on "Yes"/"No" feedback. While richer evaluative feedback could be employed in practice and converted to binary feedback, we opted for a more controlled setting to reduce confounding factors related to variations in speech content. Our contributions are as follows:

- A user study ($N = 28$) demonstrating links between prosodic features and task-related features in a Wizard-of-Oz interactive RL setting (grid-world task).
- A pilot study (30 mins of audio from a single teacher) suggesting a similar role of prosody in speech-augmented demonstrations for three Atari games.
- Preliminary evidence that incorporating prosody in interactive learning algorithms can improve learning performance.
- Documented open source data collection pipelines, including a visualization replay tool, to facilitate similar data collection efforts by other researchers.[1]

2 RELATED WORK

We provide a brief overview of work on use of human speech to augment two learning paradigms evaluated in this work: (a) interactive RL (Sec. 2.1), and (b) learning from demonstration (Sec. 2.2).

2.1 Speech-assisted Interactive RL

Some prior works which leverage human feedback during reinforcement learning tasks [54] do so via voice [16, 17, 20, 21, 47]. Tenorio et al. [47] perform reward shaping using SARSA [35, 43, 46] with human voice. Under their setup, the voice-based feedback is provided as the robot is executing the task. However, rewards are predefined for certain words in a vocabulary of 250 words such as +50 for "excellent", −50 for "terrible" etc., and no prosodic information is used. Krenig et al. [20] train RL agents with action advice in the form of human voice, such that a set of predefined words directly map to an underlying action from a discrete action set. Krenig et al. [21] use sentiment analysis to filter explanations into advice of what to do and warnings of what to avoid. Nicolescu et al. [31] demonstrated the role of verbal cues both during demonstrations and as feedback from the human teacher during the agent's learning process, to facilitate learning of navigation behaviors on a mobile robot, with a limited vocabulary of words to indicate relevant parts of the workspace or actions that a robot must execute. However, all of these prior works focus on the spoken words and do not leverage prosody from human speech—an informative and rich signal of human intent which has the potential to further enhance learning [37].

Kim et al. [17] use affective human speech feedback over 25 msec audio snippets to improve a social waving behavior using Q learning. They use three prosody features (total band energy, variance of log-magnitude-spectrum, variance of log-spectral-energy) to learn the wave that optimally satisfies a human tutor. We build on this work to further understand how prosodic features relate to RL-specific features, such as the reward and advantage function, with the goal to inform the design of future algorithms with an underlying RL formulation (e.g., interactive RL, imitation learning) which can be more sample efficient by leveraging prosodic information.

2.2 Speech-assisted Learning from Demonstration

Prior research in learning from demonstrations (LfD) [2, 33] has utilized human speech signals accompanying demonstrations. Nicolescu and Mataric [31] demonstrate the role of verbal cues both during demonstration and feedback from the human teacher to facilitate learning or sequential arrangement of navigation behaviors on a mobile robot. However, they use a fixed vocabulary of words that a demonstrator can use to indicate relevant parts of the workspace or actions that a robot must execute. Pardowitz et al. [34] show how vocal comments with a demonstration can augment subtask similarity and learning the task model (task precedence graph) for a simple table setting task. They use a fixed set of seven vocal comments which are mapped one-to-one with features relevant to the task. Rybski et al. [36] learn a planning model from human demonstrations and dialog for a mobile robot. They map human utterances to match them against a set of predefined natural language commands and manually ground them to locations on a map which the robot has access to for learning and executing navigation behaviors. In our work, however, we study how prosody in speech relates to RL-specific features and accordingly leverage prosody to aid two interactive learning paradigms.

Recently, Saran et al. [37] characterize unrestricted speech from human teachers demonstrating multi-step manipulation tasks to a situated robot. They report differentiating properties of speech in terms of duration and expressiveness, highlighting that human prosody carries rich information useful for enhancing LfD. In our work, we propose a novel algorithm that can leverage prosody for LfD and validate it in three simulated game environments.

3 STUDY 1: VOICE FEEDBACK IN INTERACTIVE RL

This study involved analyzing voice feedback provided by human trainers in an interactive RL (intRL) setup. We investigated whether prosodic features in "Yes"/"No" feedback correlated with task performance metrics, thus shedding light on the role of prosody in implicit teaching signals.

3.1 Mixed-participant Wizard-of-Oz setup

In order to develop an algorithm that leverages implicit information in prosody, we need to first understand how people use prosody as a teaching signal. On the other hand, in order to understand prosodic behavior in this context, we need to have an algorithm that incorporates prosody in its learning, which we don't have yet. To solve this paradox, we opted for a mixed participant Wizard-of-Oz (WoZ) approach where one participant plays the role of a teacher, and the other participant with no information about the task plays the role of a Wizard. This setup makes sure that the

[1] Link to code repository for both experiments: https://github.com/sahiljain11/audio_rl

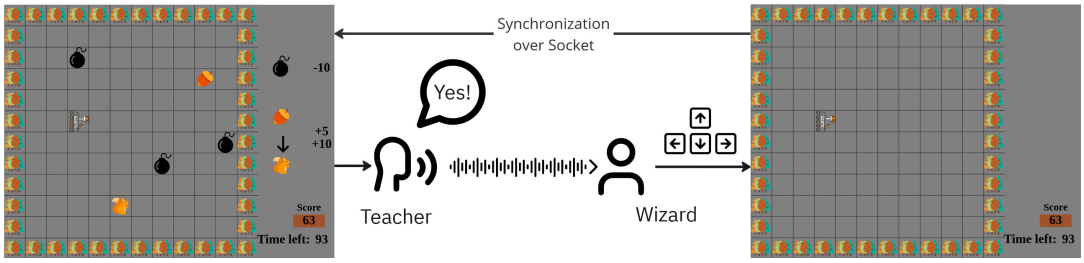


Figure 1: Mixed-participant remote Wizard-of-Oz setup with the teacher view (left) and Wizard view (right).

teacher audio we get is as close a possible to what we would expect in our target context. This approach is superior to using a baseline algorithm for the agent (e.g., intRL based on speech only) for two reasons. First, we expect the teacher to adapt to the learner, thereby potentially suppressing their prosodic signals (which was confirmed in some early pilots we ran with fixed agent trajectories). Second, the wizard's keystrokes provide us with valuable data that can be used in future research to better understand local and global interpretations of teacher feedback, independently of how well the teacher is able to teach.

Our contributed web-based WoZ interface is shown in Fig. 1. While the teacher sees the full environment and provides online verbal feedback to the agent in real-time, the Wizard is only shown a sanitized view of the environment that solely shows the grid. The Wizard also receives the teacher audio in real-time (streamed through Twilio, a secure web service) and needs to control the agent through keystrokes in response to current and past feedback from the teacher. The two environments are synchronized over web sockets to ensure consistent agent positions on both interfaces, and data is automatically and securely stored on a cloud bucket. The code for this pipeline, including a data visualization replay tool, is made open-source in order to facilitate further research in the interactive learning community[2].

3.2 Study Design

3.2.1 Participants. We recruited 28 participants to pair up in a total of 14 sessions. Most of the participants were students, except for an operations manager, a psychologist, a teacher, and a student assistant for teacher professionalization training. The mean age of the participants was 24 years old, with 16 identifying as female, seven as male, and none as other genders.

3.2.2 Experimental Setup. Our experiment used our mixed- participant WoZ setup on a Robotaxi environment [9] in which the agent had to pick up a nut and deliver it to a squirrel while avoiding three bombs (see Fig.1). Even though the agent was human-controlled in this experiment, in order to quantify task-related metrics, we modeled the underlying task as a Markov Decision Process (MDP) where discrete states represented the agent's position, actions up/ down/ left/ right were available to the agent, with deterministic state transitions and rewards at special states (shown in Fig. 1) in addition to a cost-to-live of 1 per time step. After some piloting iterations, we chose a timestep duration of 1.25 secs which made it not too boring nor too challenging for participants, while mitigating network delays (which were on average below that number).

The map was created with wall borders around the playable area. The robot location was initialized randomly at the beginning of the game, with a random initial travel direction that ensured that it could move at least three spaces in its starting direction without hitting a wall. Game elements were placed with the following constraints: bombs within at least four spaces of the robot's initial direction of travel and a Manhattan distance of at least 4 between all pairs of elements. The wizard remotely controlled the robot with arrow keys on the keyboard. In the absence of wizard input for at least one timestep, the robot started exploring the map randomly, mimicking exploration/exploitation phases typical of RL algorithms [46]. The experiment consisted of one practice round (until the goal was reached) and three game rounds for analysis.

3.2.3 Procedure. The study was approved by university ethical committees of two of our affiliations. Pairs of participants were appointed simultaneously for the study and welcomed by the examiner separately in either the teacher or the wizard role. All participants received a relevant consent form prior to the session and were compensated with a 10 EUR/USD gift card for participating in the study. All experimental sessions lasted between 15-30 minutes. Some participants optionally consented to make their data (including audio recordings) publicly available. One such sample session is included as a video recording in supplementary material.

Teacher – To establish a baseline of the teacher's voice prosody, the participant read a small paragraph given to them at the beginning of the study session (~ 30 seconds to read). During the rest of the study, the teacher was only allowed to use the words 'yes' and 'no' as feedback to the agent. To elicit richer prosody variations, we instructed the teachers to say these words as if they would to a 2 year old child. Before the study session, the teacher was told that the agent would be listening to their voice, including "how" they spoke, and acting accordingly. In reality, another participant in the role of wizard was controlling the agent in response to the teacher's verbal feedback (Fig. 1).

Wizard — The wizard was briefed on the setup of the study. They had, however, no knowledge about what the agent's task entailed or where the special states were located. Their keyboard interactions were recorded alongside other details about the game such as immediate rewards, timestamps, movements of the agent, and score. After the experiment, the wizard was asked not to talk about the experiment procedure with other potential participants.

[2]https://github.com/sahiljain11/audio_rl

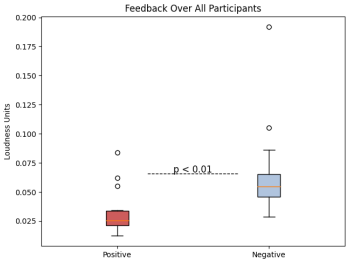


Figure 2: Yes/No balance for loudness

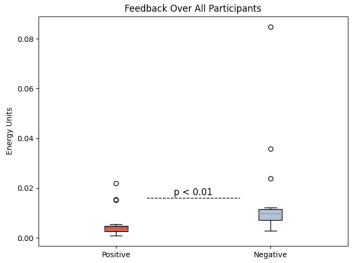


Figure 3: Yes/No balance for energy

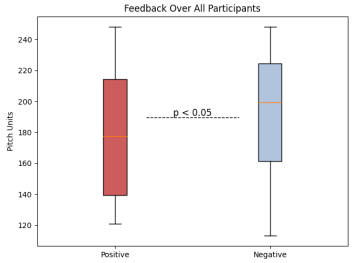


Figure 4: Yes/No balance for pitch

3.2.4 Measures. In this paper, we only focus on measures related to the teacher's data, namely:

Prosodic Features: We computed several acoustic features over the detected utterances to characterize prosody from human verbal feedback. We studied five different features to capture prosody – **utterance duration**, **utterance repetition**, **pitch**, **energy**, and **loudness**. These acoustic features have been shown to enhance semantic parsing [49], understand speech recognition failures in dialogue systems [15], and are widely used for applications to human-robot interaction [16, 42] and speech recognition [51] communities. These features were extracted from the audio recording of the experiment sessions.

MDP Features: As in most reinforcement learning settings, we model the learning task as a Markov Decision Process (MDP). The Q-value $Q(s, a)$ represents the expected utility of action a at state s if the agent follows the optimal policy. From it, the **advantage function** is calculated as: $A(s, a) = Q(s, a) - V(s)$. Recent work by Cui et al. [9] states that advantage might be a better task statistic to consider than reward for analyzing social signals (interpreted as evaluative feedback) in interactive reinforcement learning setups. Our hypothesis is that the advantage function, as a measure of relative performance of a given action at a given state, given the overall optimal policy, would significantly correlate with our prosodic features, with potential differences across different teachers based on their expressivity levels.

Teaching performance: We used agent/wizard performance as a proxy for teaching performance by counting the number of timesteps to reach the goal. Since the environment was differently set up for each participant, we normalized the performance by dividing the absolute steps of the robot by the optimal number of steps.

Demographic measures: We collected a number of demographic measures, including experience training/teaching others in a professional setting (5-point Likert item), and having a pet. The complete survey can be accessed online. [3]

3.2.5 Data processing. Transcriptions of the recordings were created using Google Cloud's Speech-To-Text [1]. The transcriptions were used to filter out silent parts and speech other than "yes" and "no" (which mostly consisted of non-verbal sounds). We encoded the positive feedback "yes" to 1 and the negative feedback "no" to −1. The prosodic feature values for each feedback utterance were calculated as follows. *Utterance duration* was estimated by calculating the time difference between the start and end timestamps of the word transcription. *Utterance repetition* was identified by analyzing word chunks within the transcription and assessing if a word was repeated consecutively as a flag $\in \{0, 1\}$. If a word was part of a repetitive word chunk, each word of the chunk got the same label. For the other prosodic features, we vectorized audio recordings with the python library Librosa [30]. To estimate the *pitch*, we employed the Librosa Yin function [30], which provides a fundamental frequency estimation. *Energy* and *loudness* were computed as $energy = \frac{\sum_1^N (x_i)^2}{N}$ and $loudness = \frac{\sum_1^N abs(x_i)}{N}$, where $x_1, ...x_N$ denotes the acoustic signal [18]. The prosodic feature *loudness* here corresponds to the sound pressure recorded by the microphone. We chose this representation since it resembles how the agent receives the audio signal, as opposed to other measures or loudness (e.g., in phon) which factor in subjective human hearing. The range of all prosodic feature values except repetition was $x \in [0, \infty)$. We combined the binary speech mapping with the prosodic features by multiplying them with the feedback values. Consequently, the sign of the prosodic values gets flipped for the negative feedback.

Due to the remote nature of the experimental setup, some timestamps would occasionally be skipped by the logger. However, in most cases it was easy to interpolate the robot's location and actions based on the previous and following timestamp for the advantage function calculation. Any missing data that was more than a couple timestamps was omitted from the analysis. Four out of 14 sessions were intact and did not need any interpolation.

3.3 Results

3.3.1 Positive/negative bias. In line with previous work [48], we observed significantly more positive than negative feedback for five out of 14 trainers (p-values ≤ 0.003 on a Bonferroni corrected chi-square goodness of fit test). The other participants did not show statistical significance. A population level chi-square goodness of fit test showed that there was significantly more positive than negative feedback over all participants ($p \leq 0.001$). Therefore, we conclude that the usage of positive and negative feedback was unbalanced and more positive feedback was used.

Our results also showed that the extent to which prosody features (loudness, pitch, and energy) were used is higher for negative than positive feedback. We did not observe the same effect for duration and repetition features. We ran t-test analyses for each prosodic feature and each participant to compare the prosodic expressions. The results can be found in Fig. 2, 3, 4. The plots show the aggregated values for each prosodic feature over all participants. The

[3] Survey URL: https://tinyurl.com/SurveyAudioRL

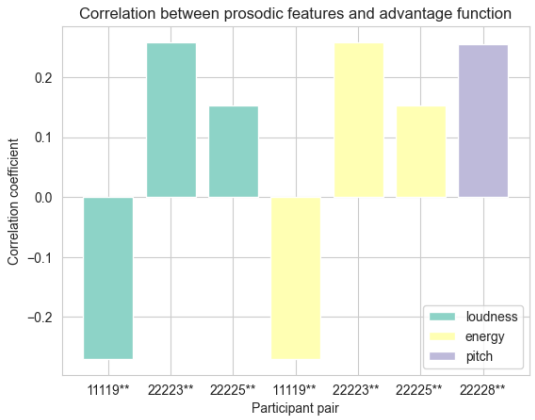


Figure 5: Statistically significant correlations between the prosodic features and the advantage values

aggregated p-values were determined with a Bonferroni correction. When we look at each prosodic feature for different participants, energy and loudness have 13 out of 14 significant differences between the prosodic features of the feedback word "yes" and "no". This highlights that participants spoke louder and with higher energy when they gave negative feedback to the agent. Pitch has a less clear distinction between positive and negative feedback. Seven out of the 14 trainers had a significantly different pitch value when saying the word "no" than "yes".

3.3.2 Link to MDP metrics. For each participant, advantage values were correlated with the corresponding word's prosodic feature. Spearman's rank correlation coefficient was used with a Bonferroni correction to determine the relation between the variables since the advantage function is not normally distributed. For the "repetition" feature, a point-biserial correlation was used.

For loudness and energy, three out of 14 correlations were significant, with a correlation coefficient of 0.15 and 0.25. This suggests that people give positive feedback with higher energy and loudness if the taken action was the best possible one. In contrast, a suboptimal action is associated with negative feedback rich in energy and loudness. One correlation had a negative coefficient of −0.25, suggesting that that person gave prosody-rich positive feedback when actions were suboptimal and prosody-rich negative feedback when the action was optimal. In order to compare the correlation results at the population level, we conducted a repeated measure correlation. The result showed that the correlation did not hold at the population level with a p-value = 0.23.

The correlation between pitch and advantage values was significant for one out of fifteen participants, with a correlation coefficient of 0.25. As mentioned above, this suggests that people give positive feedback with a higher pitch if the taken action was the best possible one. Additionally, we conducted a repeated measure correlation. The result showed that the correlation did not hold at the population level ($p = 0.51$). Figure 5 shows statistically significant correlations.

The correlations for repetition and duration were both not significant. Thus, neither of the two prosodic features was associated with the advantage values.

3.3.3 Effect of demographics. Additionally, we investigated whether training experience can be associated with being a good agent trainer. We took the item "To what extent does your profession involve teaching or training other people?" and correlated the Likert-scale answers ranging from 1 (None at all) to 5 (a great deal) with the performance metric. The correlation was statistically significant ($r = -0.68$, $p = 0.02$). Thus, having a profession that involves teaching can be associated with being a good trainer. We did not find any statistically significant associations between training competence and other demographic variables such as having a pet.

3.4 Algorithmic Implications

Building on our empirical results, we tested the applicability of a prosody-sensitive learning paradigm by incorporating the combined speech and prosody feedback into a human reward function of an intRL agent. Since previous studies have shown promising results with intRL agents trained by using voice feedback [7] and prosodic feedback [18], we built on this research by demonstrating that added prosody offers additional benefits beyond explicit voice feedback.

We incorporated prosody in an interactive RL algorithm and tested it in an offline manner on a subset of the sessions collected in our experiment (only the four sessions were no interpolation was needed). The intRL algorithm we chose was TAMER [19] due to its popularity, and compatibility with many different feedback modalities [8, 19, 29]. For implementation details, see Appendix A.3.

Our implemented intRL algorithm took combined explicit and implicit voice input as human feedback to learn the feedback function H. We implemented a prosody-augmented version of TAMER while accounting for individual variations by normalizing it with respect to each individual's baseline mean and standard deviation (z-standardization). The feedback function was determined by taking the mean of all three prosodic feature values for both positive and negative feedback. We decided to make one combined prosodic feature value based on the finding that the features are highly correlated, which implies that they are used together.

The performance of the algorithm was evaluated by assessing the learned policy of H. For this, all possible optimal actions for each state in the environment were determined. The policy for H was determined by choosing each action using a greedy approach. The absolute number of optimal actions was counted and used as the evaluation metric.

In Fig. 6, the performances are visually displayed across the four participants. At the population level, although not statistically significant due to small sample size, the prosody condition had the highest mean value of optimal actions. However, when looking at individual participants, for three out of the four participants, the prosody condition statistically significantly outperformed the TAMER baseline. These preliminary results are promising and will be followed up on in larger studies in the future.

4 STUDY 2: PROSODY IN AUDIO-AUGMENTED DEMONSTRATIONS

In this study, we collected audio data from a single demonstrator playing three Atari games while providing only "yes" and "no" utterances. The demonstrator's prosodic features were analyzed

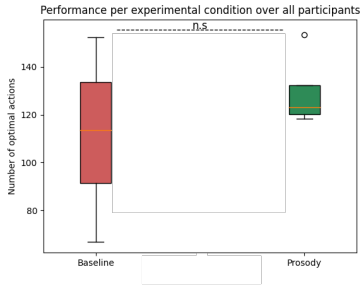


Figure 6: Performance of prosody-augmented TAMER (i.e., speech content + prosody) versus TAMER baseline (i.e., only speech content) in an offline RL setting ($N = 4$)).

(a) Ms. Pacman (b) Seaquest (c) Space Invaders

Figure 7: Atari games used in our study to understand a human demonstrator's speech patterns.

alongside game dynamics to understand the relationship between prosody and task performance. Particularly, by restricting the usage of words ("yes" and "no" only) and working with a single demonstrator, we isolate the impact of speech prosody from spoken words and user differences respectively.

4.1 Study Design

4.1.1 Experimental Setup. We collect demonstration and audio data from a single demonstrator using a customized simulation interface for three Atari Games shown in Figure 7: (1) Ms. Pac-Man, (2) Seaquest, (3) Space Invaders. We use these games due to the diversity in their objectives and reward schemes. The demonstrator provides state-action data (states are images of the game screen, and actions are keystrokes used for game-play) and audio data in the form of only "yes" and "no" utterances (collected over the web via the demonstrator's device microphone) for 10 minutes of gameplay per game. This data was collected remotely through a web interface and recorded on a remote server. Screenshots of the data collection interface are shown in Appendix B.2. We asked the user to demonstrate examples of how to play the games to the best of their ability by using their keyboard as well as by using their microphone so the Atari agent will learn how to play the game by observing both their keyboard strokes as well as their voice. We also stated that the character will move and be controlled by their keystrokes only, but the agent can understand the words "yes" and "no", and can identify the pitch or intensity with which they say these words.

4.1.2 Procedure. The user was provided an opportunity to practice the game before their demonstrations were recorded to help them better understand the dynamics and reward structure of the game. The user was instructed to play the games in a quiet environment to minimize the possibility of any background noise being recorded. Data is collected using the Atari Grand Challenge (AGC) interface [22]. However, an additional functionality is added to record the human audio synchronized with demonstrated states and actions (Appendix B.2). The start and end times of each utterance are detected with Google's speech-to-text API, and the reward values are provided by the logs of the AGC interface.

4.2 Measures

Prosodic Features: After collecting and synchronizing the state, action, and speech data, prosodic feature values for each utterance were computed by loading the data in the python library Librosa and a default sampling rate of 22050 [30]. We use *utterance duration*, *utterance repetition*, *energy*, *pitch*, and *loudness* as the different prosodic features based on prior work in the speech recognition and learning from demonstration communities [15, 37]. Following these, we report the mean and maximum values for pitch, loudness, and energy for each speech utterance. We also report the total energy (cumulative energy sum per utterance).

MDP Features: In prior work studying unrestricted speech prosody for robot learning by Saran et al. [37], demonstration rewards or errors are shown to be the most promising MDP statistic to leverage with prosody to enhance learning. Thus, we study how often the demonstrator uses "yes" and "no" utterances, as well as the connection of prosody with the underlying ground truth rewards in the game.

Based on prior studies that have identified that prosodic features can convey semantic meaning for words used in speech [32], we expect that the prosodic feature values can help identify the nature of speech used (positive/negative) by being significantly different for the "yes" and "no" utterances. The meaning or nature of speech used is also evaluated against the underlying ground truth rewards from the game design.

4.3 Results

4.3.1 Positive/Negative bias: From the playback of synchronized audio along with demonstrations, we observe that the demonstrator uses utterances as both a reaction to recent past events and anticipatory guidance for future events during a demonstration. For this reason, we use a 0.5 second buffer before the start time and after the end time of each audio utterance when computing features or reward values accompanying an utterance. The average duration of utterances is 0.76 seconds (Ms. Pacman), 0.61 seconds (Seaquest), and 0.75 seconds (Space Invaders) (Fig. 13).

We compute the overall duration of "yes" and "no" utterances compared to the total duration of a demonstration (see Fig. 14 in Appendix B.1). We find that the demonstrator uses significantly more "yes" utterances compared to "no" utterances (Fig. 14) for all three games ($p < 0.05$). The proportion of "yes" utterances is the highest for Ms. Pacman which is a multi-objective game where the agent has to move rapidly by escaping ghosts and procuring food palettes. There is less room for respite when the game begins,

which might have made the demonstrator more active with their speech in this game compared to other games. Thus, "no" utterances or negative feedback are used for rare events during the games, where the agent loses points or dies (end of episode), whereas "yes" or positive feedback is used more frequently to indicate good progression in the game.

Next, we compute prosodic feature values accompanying the "yes" and "no" utterances for each game (see Fig. 16 in Appendix B.1). We find that mean energy, maximum energy, total energy, mean loudness, and maximum loudness are significantly higher ($p < 0.05$) for negative feedback ("no") compared to positive feedback ("yes") for all the games (except maximum loudness during Space Invaders). This finding is similar to that of Sec. 3.3.1 for interactive RL.

4.3.2 Link to MDP statistics: The cumulative sum of rewards for trajectory snippets accompanying speech utterances is significantly higher for "yes" utterances versus "no" utterances (Fig. 15). This finding intuitively states that the cumulative reward during positive feedback ("yes") is higher than the cumulative reward during negative feedback ("no") from the demonstrator.

Table 1: Spearman correlation between mean pitch and ground truth rewards for "yes" and "no" utterances used along with demonstrations to three different Atari games.

	Yes	No
Ms. Pacman	0.13	−0.12
Seaquest	0.2^{*}	-0.77^{*}
Space Invaders	0.37^{**}	-0.38^{*}

While mean pitch is not significantly different for "yes" and "no" utterances (Fig. 16), we find another pattern for mean pitch (but not for other prosodic features)— there is a positive spearman correlation between mean pitch and cumulative sum of rewards for trajectory snippets accompanying "yes" audio utterances (Table 1). Similarly, we also observe a negative spearman correlation between mean pitch and cumulative rewards for trajectory snippets accompanying "no" audio utterances. This finding reveals that "yes" utterances with higher pitch have higher ground truth returns associated with them and "no" utterances with higher pitch have lower returns associated with them, i.e. the more the gain or loss with an action during the demonstration, the more emphatically the corresponding word is spoken. We do not find a consistent pattern of correlation for any other prosodic features such as energy, loudness. Thus, while pitch might not distinguish between the yes and no utterances, it can correlate with the magnitude of reward values based on the type of speech usage (positive/negative). These results hint towards prosody in speech revealing information about the underlying rewards, thereby potentially enhancing the sample efficiency for LfD methods which first learn the underlying reward function to train the agent policy (inverse reinforcement learning).

4.4 Algorithmic Implications

While several recent works [12, 24, 52] have utilized object or environmental audio for learning from demonstration, to the best of our knowledge, spoken and prosodic cues of human teachers have not been leveraged with deep LfD algorithms. Based on the findings in Sec. 4.3, we propose an efficient technique to leverage speech in the form of an auxiliary loss for reward learning. The auxiliary loss leverages both spoken words and prosodic features from human speech to guide the training of a deep inverse reinforcement learning method T-REX [4] for the Atari game-playing domain. T-REX trains a deep reward model by comparing the performance of pairs of demonstrated trajectory snippets. The final agent policy is trained via an RL algorithm (such as PPO [41]) on the learned reward. Since the only two words used by the demonstrator signify if an event during the demonstration is positive (yes) or negative (no) as shown in Sec. 4.3 and Fig. 15 in Appendix B.1, the "yes" and "no" labels can create contrasting categories of sample trajectory snippets for reward learning. The reward values predicted by the reward network can be compared for these contrasting trajectory snippets, and in turn the reward network's parameters can be appropriately penalized during training. We use the prosodic feature values to determine how to scale the similarity between pairs of snippets accompanying audio utterances. Based on the finding in Sec. 4.3 and Table 1, correlation of mean pitch with reward values motivate us to scale the similarity between two demonstration snippets according to the difference in their corresponding pitch.

4.4.1 Details of Model Training. The audio, state, and action data are synchronized post data collection to accurately extract demonstrated trajectory snippets accompanied by audio utterances for training the T-REX agent [4]. We use a contrastive loss to guide the training of the reward network for T-REX [4] which we call the contrastive audio loss (CAL). Given a sequence of m demonstrations ranked from worst to best, $\tau_1, \ldots, \tau_m$, a parameterized reward network $\hat{r}_\theta$ is trained with a cross-entropy loss over a pair of trajectories ($\tau_i \prec \tau_j$), where τ_j is ranked higher than τ_i. We add CAL as an auxiliary loss for training the reward network with additional trajectory pairs τ_m and τ_n which are accompanied by audio utterances provided by the demonstrator, so the new loss function becomes:

$$\mathcal{L}(\theta) = -\sum_{\tau_i \prec \tau_j} \log \frac{\exp \sum_{s\in\tau_j} \hat{r}_\theta(s)}{\exp \sum_{s\in\tau_i} \hat{r}_\theta(s) + \exp \sum_{s\in\tau_j} \hat{r}_\theta(s)} + \alpha\left[CAL\Big(\sum_{s\in\tau_m} r^a_m(s), p_m, \sum_{s\in\tau_n} r^a_n(s), p_n\Big)\right] \tag{1}$$

where

$$CAL\Big(R^a_m, p_m, R^a_n, p_n\Big) = -w_{mn} \log \frac{\exp(\text{sim}(R^a_m, R^a_n)/t_{mn})}{\sum_{k=1}^{2N} 1_k \exp(\text{sim}(R^a_m, R^a_k)/t_{mk})} \tag{2}$$

$$t_{mn} = t_0 + |p_m - p_n| \tag{3}$$

$$\text{sim}(R^a_m, R^a_n) = \frac{1}{1 + |R^a_m - R^a_n|} \tag{4}$$

$\sum_{s\in\tau_m} r^a_m(s)$ (Equation 1) or R^a_m (Equation 2) represents the undiscounted cumulative sum of rewards corresponding to every state s from the trajectory snippet τ_m. $\text{sim}(R^a_m, R^a_n)$ in Equations 2,4 represents a similarity measure between the undiscounted cumulative sum of rewards for two trajectory snippets, and p_m in Equations 2,3 represents a scalar prosodic feature value for the audio chunk a_m accompanying τ_m. The difference in prosodic features $|p_m - p_n|$

Table 2: Average performance over 30 rollouts for three Atari game agents trained in the LfD paradigm.

	T-REX	T-REX + CAL
Ms. Pacman	414.0±14.9	**663.3±121.3**
Seaquest	679.3±14.3	**704.0±6.0**
Space Invaders	1235.2±102.3	**1781.3±127.6**

is normalized with a softmax function for all the differences from a batch. t_{mn} in Equations 2,3 is the temperature parameter, and t_0 is an offset temperature value to avoid numerical inconsistencies. Based on findings in Sec. 4.3, we use mean pitch as the prosodic feature to determine the temperature values automatically. The loss function accumulates the impact of audio over the entire trajectory snippet for each trajectory pair used as input to the network.

We evaluate the performance of each game with and without CAL augmentation during training. Performance is measured in terms of game score averaged over 30 trials for three seeds, with the highest result reported among the three seeds. Standard error is also reported along with mean performance.

4.4.2 Implications. Results for the three games are shown in Table 2, where incorporating speech cues via CAL improves average performance of each game. This highlights the efficacy of incorporating speech cues from a single demonstrator for learning from demonstration. Both the content of speech (what is being said), and prosody (how something is said) are useful in guiding agents trained via the T-REX algorithm. Here, the content of words have a direct mapping to positive and negative events when comparing a pair of trajectory snippets, whereas prosody is used in a manner that scales the similarity between two spoken utterances (and in turn the similarity in the predicted reward values) based on the magnitude of their corresponding pitch.

In summary, we find that the content of speech is more indicative of the underlying reward type (low/high), while prosodic cues are more indicative of scale of rewards. We take the findings from this study and build on them by augmenting a deep imitation learning algorithm with speech. We present an auxiliary loss (contrastive audio loss) to leverage simple predefined speech cues from a single demonstrator to train T-REX agents for high-dimensional Atari games. Our investigation shows that underutilized speech cues can effectively guide agent learning in the learning from demonstration paradigm. Leveraging the underutilized speech modality can thus enable sample efficient training and save time spent by human teachers demonstrating tasks to AI agents (extracting more information from a fixed set of demonstrations with minimal cost to record speech information and without requiring any additional time spent by the teacher).

5 DISCUSSION

Speech is a low-effort and rich source of information that humans naturally provide when teaching artificial learning agents. In addition to spoken words from natural language, speech also contains prosodic cues which can be informative towards demystifying the underlying goals or progress during the task to artificial learning agents. Our exploratory studies about the integration of prosody as a teaching signal in agent learning environments have yielded promising results, illustrating the potential of prosodic cues to enhance the efficacy of human-agent interactions in both reinforcement learning and learning from demonstration scenarios.

One of the primary limitations of our studies relates to the controlled nature of the feedback environment, where we restricted the verbal input to binary "Yes"/"No" responses. While this design choice was intended to minimize confounding variables, it also limits the richness of the feedback and may not fully represent more dynamic real-world interactions where verbal feedback can be more nuanced. The generalizability of our results across different tasks and agent embodiments (beyond simulations) also remains an open question. Our studies were conducted with a specific set of tasks, and future work should explore how well our findings would translate to other contexts or more complex decision-making environments.

To better account for variations across human teachers, future studies could incorporate adaptive learning systems that are sensitive to individual teaching styles and learner responses. Such systems can adjust their interpretation of prosodic cues based on the specific teacher-learner dyad, potentially through personalized calibration sessions or real-time feedback mechanisms. Future research could also explore the integration of richer verbal feedback and the development of more sophisticated models for prosody recognition and interpretation. The potential for real-time adaptation to the teacher's prosodic patterns offers an intriguing avenue for developing more responsive and sensitive learning agents. Moreover, incorporating multi-modal feedback, where prosodic cues are considered in conjunction with other non-verbal cues such as gaze, gestures, or facial expressions, could provide a more holistic approach to understanding and leveraging natural human communication patterns for interactive machine learning.

6 CONCLUSION

In this work, we present analyses of human prosody during two interactive learning paradigms — interactive reinforcement learning (intRL) and learning from demonstration (LfD). We find that prosody from human teachers is expressed more strongly during negative feedback compared to positive feedback in both learning paradigms. Additionally, we find correlations between prosodic feature values and the advantage function for intRL, and between prosodic feature values and the reward function for LfD. With these results, we motivate the design of novel algorithms in the two learning paradigms which leverage human prosody during learning. Our proof of concept experiments reveal that prosody is a promising modality to enhance learning and improve sample efficiency.

This paper highlights the untapped potential of prosody in enhancing agent learning from human interaction. By advocating for the integration of prosody-sensitive algorithms and providing empirical evidence of their efficacy, we aim to advance the field of human-agent interaction and pave the way for more intuitive and efficient learning paradigms.

ACKNOWLEDGMENTS

The authors would like to thank Taylor Kessler Faulkner, Andrea Thomaz, Scott Niekum for their valuable feedback, and Ojas Patel, Rakesh Johny for contributing to the codebase.

REFERENCES

[1] [n. d.]. Google Speech-to-Text API. https://cloud.google.com/speech-to-text.

[2] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems* 57, 5 (2009), 469–483.

[3] Yusuf Aytar, Tobias Pfaff, David Budden, Tom Le Paine, Ziyu Wang, and Nando de Freitas. 2018. Playing hard exploration games by watching youtube. *arXiv preprint arXiv:1805.11592* (2018).

[4] Daniel S Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. 2019. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. *arXiv preprint arXiv:1904.06387* (2019).

[5] Ze-Yin Chen, Yi-Jun Li, Miao Wang, Frank Steinicke, and Qinping Zhao. 2021. A reinforcement learning approach to redirected walking with passive haptic feedback. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 184–192.

[6] Francisco Cruz, German I Parisi, and Stefan Wermter. 2018. Multi-modal feedback for affordance-driven interactive reinforcement learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[7] Francisco Cruz, Johannes Twiefel, Sven Magg, Cornelius Weber, and Stefan Wermter. 2015. Interactive reinforcement learning through speech guidance in a domestic scenario. https://doi.org/10.1109/ijcnn.2015.7280477

[8] Yuchen Cui, Qiping Zhang, Alessandro Allievi, Peter Stone, Scott Niekum, and W. Knox. 2020. The EMPATHIC Framework for Task Learning from Implicit Human Feedback. (09 2020).

[9] Yuchen Cui, Qiping Zhang, Alessandro Allievi, Peter Stone, Scott Niekum, and W Bradley Knox. 2020. The EMPATHIC Framework for Task Learning from Implicit Human Feedback. *arXiv preprint arXiv:2009.13649* (2020).

[10] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. 2020. See, hear, explore: Curiosity via audio-visual association. *arXiv preprint arXiv:2007.03669* (2020).

[11] Taylor A Kessler Faulkner, Elaine Schaertl Short, and Andrea L Thomaz. 2020. Interactive reinforcement learning with inaccurate feedback. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7498–7504.

[12] Dhiraj Gandhi, Abhinav Gupta, and Lerrel Pinto. 2019. Swoosh! Rattle! Thump!-Actions that Sound. (2019).

[13] Anna Gergely, Anna Gábor, Márta Gácsi, Anna Kis, Kálmán Czeibert, József Topál, and Attila Andics. 2023. Dog brains are sensitive to infant-and dog-directed prosody. *Communications Biology* 6, 1 (2023), 859.

[14] Prasoon Goyal. 2022. *Using natural language to aid task specification in sequential decision making problems*. Ph. D. Dissertation.

[15] Julia Hirschberg, Diane Litman, and Marc Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech communication* 43, 1–2 (2004), 155–175.

[16] Elizabeth S Kim, Dan Leyzberg, Katherine M Tsui, and Brian Scassellati. 2009. How people talk when teaching a robot. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 23–30.

[17] Elizabeth S Kim and Brian Scassellati. 2007. Learning to refine behavior using prosodic feedback. In *2007 IEEE 6th International Conference on Development and Learning*. IEEE, 205–210.

[18] Elizabeth S. Kim and Brian Scassellati. 2007. Learning to refine behavior using prosodic feedback. (2007), 205–210. https://doi.org/10.1109/DEVLRN.2007.4354072

[19] W. Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement. *Proceedings of the fifth international conference on Knowledge capture* (9 2009). https://doi.org/10.1145/1597735.1597738

[20] Samantha Krening. 2018. Newtonian action advice: Integrating human verbal instruction with reinforcement learning. *arXiv preprint arXiv:1804.05821*.

[21] Samantha Krening, Brent Harrison, Karen M Feigh, Charles Lee Isbell, Mark Riedl, and Andrea Thomaz. 2016. Learning from explanations using sentiment and advice in RL. *IEEE Transactions on Cognitive and Developmental Systems* 9, 1 (2016), 44–55.

[22] Vitaly Kurin, Sebastian Nowozin, Katja Hofmann, Lucas Beyer, and Bastian Leibe. 2017. The atari grand challenge dataset. *arXiv preprint arXiv:1705.10998* (2017).

[23] Guangliang Li, Randy Gomez, Keisuke Nakamura, and Bo He. 2019. Human-centered reinforcement learning: A survey. *IEEE Transactions on Human-Machine Systems* 49, 4 (2019), 337–349.

[24] Hongzhuo Liang, Shuang Li, Xiaojian Ma, Norman Hendrich, Timo Gerkmann, Fuchun Sun, and Jianwei Zhang. 2019. Making sense of audio vibration for liquid height estimation in robotic pouring. *arXiv preprint arXiv:1903.00650* (2019).

[25] Jacky Liang, Fei Xia, Wenhao Yu, Andy Zeng, Montserrat Gonzalez Arenas, Maria Attarian, Maria Bauza, Matthew Bennice, Alex Bewley, Adil Dostmohamed, Chuyuan Kelly Fu, Nimrod Gileadi, Marissa Giustina, Keerthana Gopalakrishnan, Leonard Hasenclever, Jan Humplik, Jasmine Hsu, Nikhil Joshi, Ben Jyenis, Chase Kew, Sean Kirmani, Tsang-Wei Edward Lee, Kuang-Huei Lee, Assaf Hurwitz Michaely, Joss Moore, Ken Oslund, Dushyant Rao, Allen Ren, Baruch Tabanpour, Quan Vuong, Ayzaan Wahid, Ted Xiao, Ying Xu, Vincent Zhuang, Peng Xu, Erik Frey, Ken Caluwaerts, Tingnan Zhang, Brian Ichter, Jonathan Tompson, Leila Takayama, Vincent Vanhoucke, Izhak Shafran, Maja Mataric, Dorsa Sadigh, Nicolas Heess, Kanishka Rao, Nik Stewart, Jie Tan, and Carolina Parada. 2024. Learning to Learn Faster from Human Feedback with Language Model Predictive Control. (2024). arXiv:2402.11450 [cs.RO]

[26] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. 2020. A review on interactive reinforcement learning from human social feedback. *IEEE Access* 8 (2020), 120757–120765.

[27] Jinying Lin, Qilei Zhang, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. 2020. Human social feedback for efficient interactive reinforcement agent learning. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 706–712.

[28] Andrea Lockerd and Cynthia Breazeal. 2004. Tutelage and socially guided robot learning. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vol. 4. IEEE, 3475–3480.

[29] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. 2017. Interactive Learning from Policy-Dependent Human Feedback. (2017), 2285–2294.

[30] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8.

[31] Monica N Nicolescu and Maja J Mataric. 2003. Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. 241–248.

[32] Lynne C Nygaard, Debora S Herold, and Laura L Namy. 2009. The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive science* 33, 1 (2009), 127–146.

[33] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. 2018. An algorithmic perspective on imitation learning. *arXiv preprint arXiv:1811.06711* (2018).

[34] Michael Pardowitz, Steffen Knoop, Ruediger Dillmann, and Raoul D Zollner. 2007. Incremental learning of tasks from user demonstrations, past experiences, and vocal comments. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 37, 2 (2007), 322–332.

[35] Gavin A Rummery and Mahesan Niranjan. 1994. *On-line Q-learning using connectionist systems*. Vol. 37. University of Cambridge, Department of Engineering Cambridge, UK.

[36] Paul E Rybski, Kevin Yoon, Jeremy Stolarz, and Manuela M Veloso. 2007. Interactive robot task training through dialog and demonstration. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. ACM, 49–56.

[37] Akanksha Saran, Kush Desai, Mai Lee Chang, Rudolf Lioutikov, Andrea Thomaz, and Scott Niekum. 2022. Understanding acoustic patterns of human teachers demonstrating manipulation tasks to robots. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.

[38] Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. 2018. Human gaze following for human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8615–8621.

[39] Akanksha Saran, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. 2020. Understanding teacher gaze patterns for robot learning. In *Conference on Robot Learning*. PMLR, 1247–1258.

[40] Akanksha Saran, Ruohan Zhang, Elaine Schaertl Short, and Scott Niekum. 2021. Efficiently Guiding Imitation Learning Agents with Human Gaze. *International Conference on Autonomous Agents and Multiagent Systems* (2021).

[41] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[42] Elaine Schaertl Short, Mai Lee Chang, and Andrea Thomaz. 2018. Detecting contingency for HRI in open-world environments. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 425–433.

[43] Satinder P Singh and Richard S Sutton. 1996. Reinforcement learning with replacing eligibility traces. *Machine learning* 22, 1-3 (1996), 123–158.

[44] Maria Spinelli, Mirco Fasolo, and Judi Mesman. 2017. Does prosody make the difference? A meta-analysis on relations between prosodic aspects of infant-directed speech and infant outcomes. *Developmental Review* 44 (2017), 1–18.

[45] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[46] Richard S Sutton, Andrew G Barto, et al. 1998. *Introduction to reinforcement learning*. Vol. 135. MIT press Cambridge.

[47] Ana C Tenorio-Gonzalez, Eduardo F Morales, and Luis Villaseñor-Pineda. 2010. Dynamic reward shaping: training a robot by voice. In *Ibero-American conference on artificial intelligence*. Springer, 483–492.

[48] Andrea Thomaz and Cynthia Breazeal. 2006. Reinforcement Learning with Human Teachers: Evidence of Feedback and Guidance with Implications for Learning Performance. *Proceedings of the National Conference on Artificial Intelligence* 1 (01 2006).

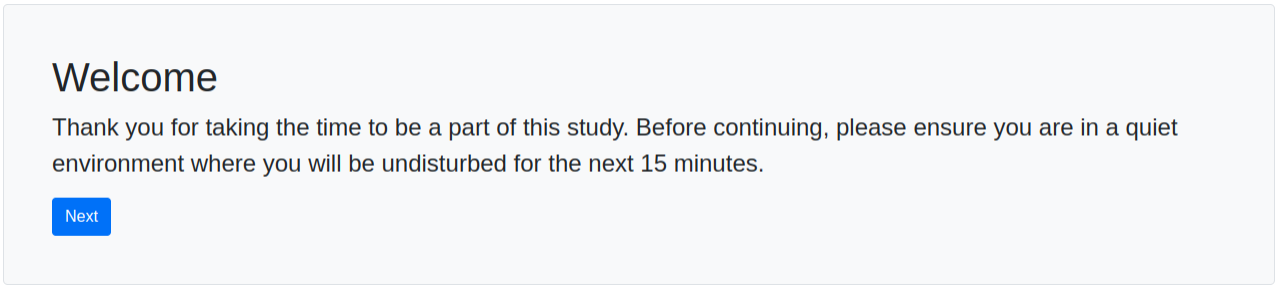
[49] Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf. 2017. Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information. *arXiv preprint arXiv:1704.07287* (2017).

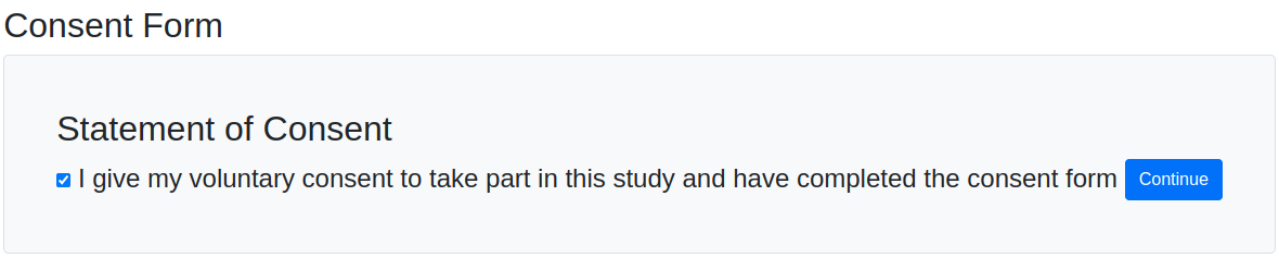
[50] Federico Ghelli Visi and Atau Tanaka. 2020. Towards assisted interactive machine learning: exploring gesture-sound mappings using reinforcement learning. In *ICLI 2020—the fifth international conference on live interfaces*. 9–11.
[51] Dong Yu and Li Deng. 2016. *AUTOMATIC SPEECH RECOGNITION*. Springer.
[52] Kevin Zhang, Mohit Sharma, Manuela Veloso, and Oliver Kroemer. 2019. Leveraging Multimodal Haptic Sensory Data for Robust Cutting. In *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*. IEEE, 409–416.
[53] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. 2020. Human gaze assisted artificial intelligence: A review. In *IJCAI: Proceedings of the Conference*, Vol. 2020. NIH Public Access, 4951.
[54] Ruohan Zhang, Faraz Torabi, Lin Guan, Dana H Ballard, and Peter Stone. 2019. Leveraging human guidance for deep reinforcement learning tasks. *arXiv preprint arXiv:1909.09906* (2019).
[55] Ruohan Zhang, Faraz Torabi, Garrett Warnell, and Peter Stone. 2021. Recent advances in leveraging human guidance for sequential decision-making tasks. *Autonomous Agents and Multi-Agent Systems* 35, 2 (2021), 1–39.

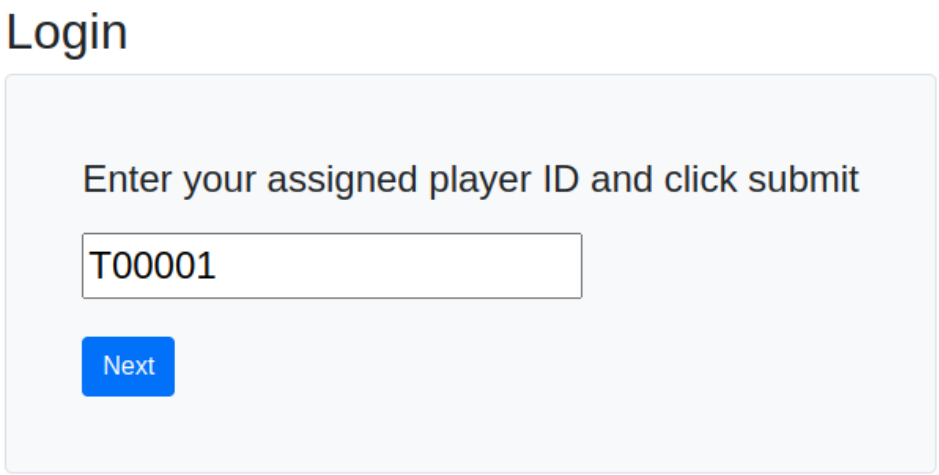
A APPENDIX: INTERACTIVE RL

A.1 Data Collection Interface (Teacher)

Below we share screenshots of the data collection interface provided to the teacher during a user study session.







Audio Recording

First, we will establish a baseline of how you speak.

If the browser prompts for microphone access, click "Allow". When you are ready, click "Start Recording" and speak the paragraph below in a normal voice. When finished, press the button again to stop the recording.

Read the paragraph below:

The beige hue on the waters of the loch impressed all, including the French queen, before she heard that symphony again, just as young Arthur wanted. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache.

Stop Recording

Instructions

What is this study about?

As a participant of this study, you will teach a robot how to solve a game. In order to teach the agent how to win a game, you will provide the agent with negative or positive feedback. This feedback will help the agent understand when it's doing a good or bad job.

Please click "Next" when you're ready to move on.

Next

Instructions

Game Instructions

The goal of the game is to deliver the acorn to the squirrel while avoiding the bombs. That is, the robot must drive over the acorn and then over the squirrel. Driving over a bomb will incur a penalty and reset the state of the game. The robot can move up, down, left, or right.

Round 0 will be a practice round. After that, rounds 1-3 will use the same map but a different starting location.

Point System

You will start with 100 points. Successfully picking up the acorn will give you 5 points and feeding the squirrel will give you 10 extra points. However, if you run into a bomb, the game starts over and you lose 10 points.

YOUR JOB

Your job is to use your voice to help the computer win the game. Please **ONLY** say `"Yes" or "No"` to provide feedback. Note that the computer is sensitive to how say these words, i.e., the expressivity of your voice, so imagine you are talking to a 2 year old or to a pet.

Keep in mind the computer cannot see the acorn, squirrels, or bombs, so it only has your voice for guidance.

Try to help the computer complete each round as quickly as possible.

Instructions

Key Points

- You can ONLY say “yes” and “no”, no other words, but the robot is sensitive to how you say these words. Make sure you are as expressive as possible. Pretend you are talking to a 2 year old, that’s about how socially receptive the computer is!
- The computer cannot see the nut, squirrel, or bombs, only an empty grid! It will only be able to learn from your audio, so make sure you give it clear cues when it does a good thing versus a bad thing.
- Do not use the back button in the browser. Let the study coordinator know if anything goes wrong.

Please click "Start Practice" when you're ready to start the game.

Start Practice

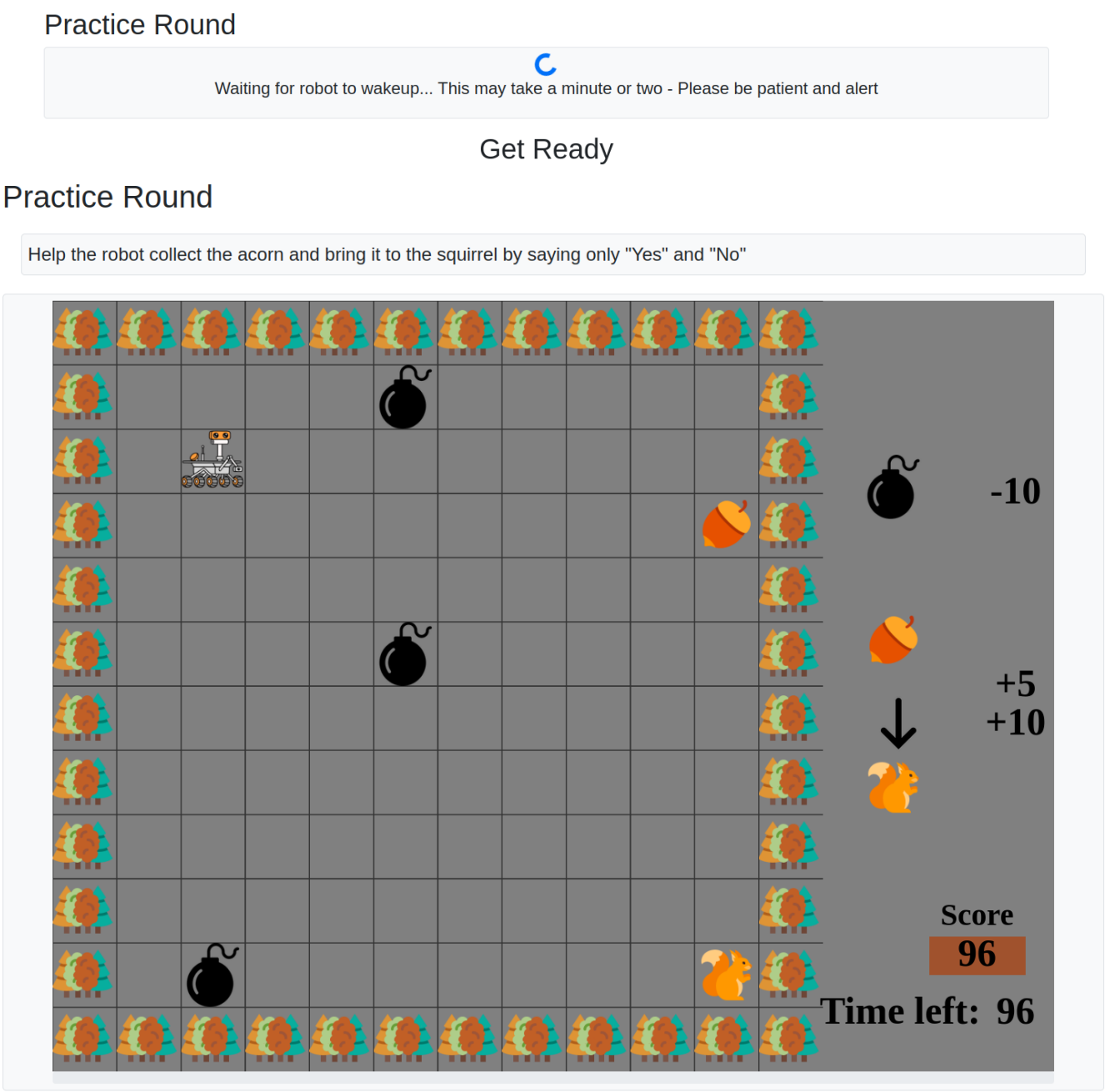


Figure 8: The teacher is able to see all game objects and use their speech to provide feedback for both the trial and actual games. However, they are unable to control the agent

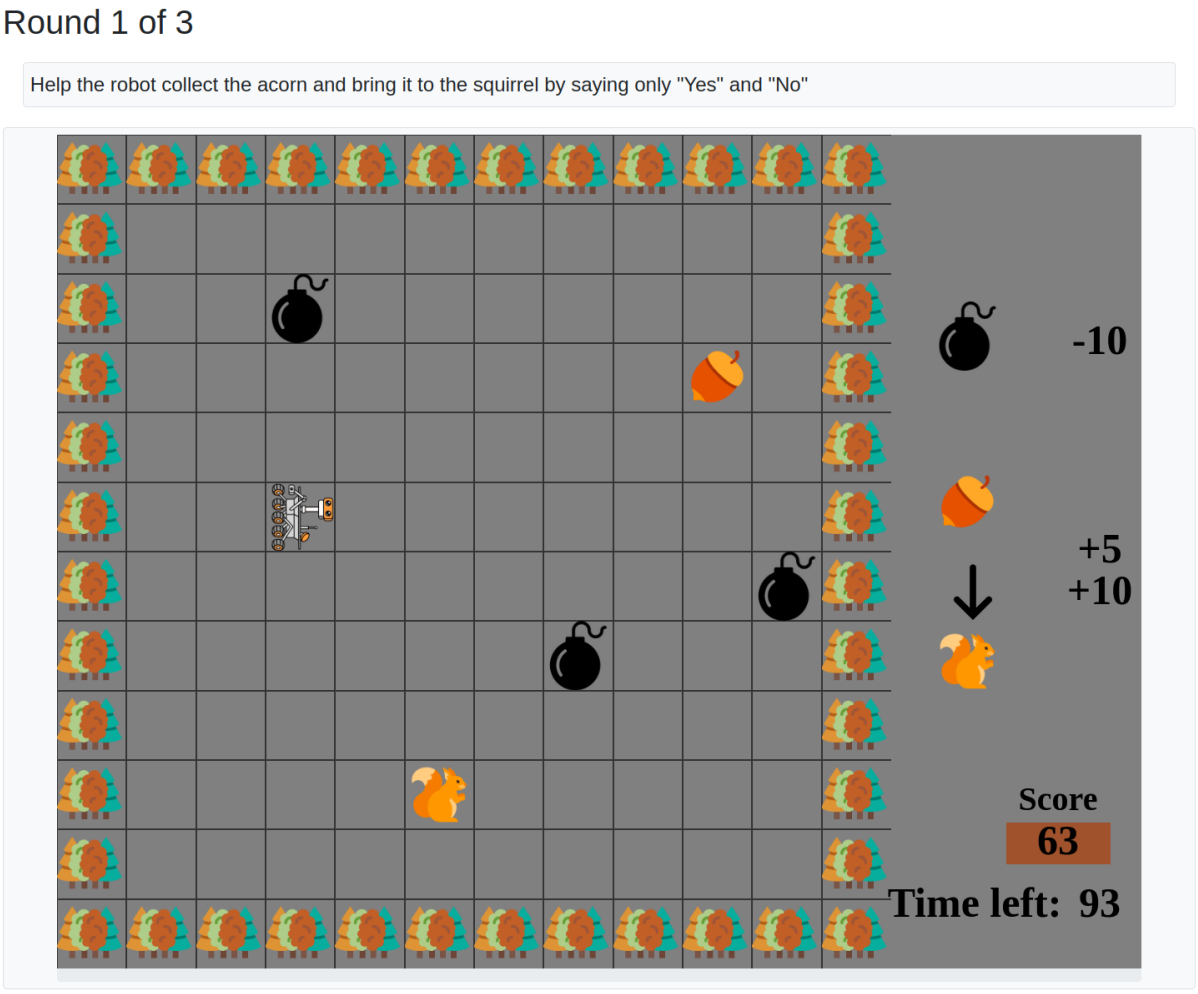
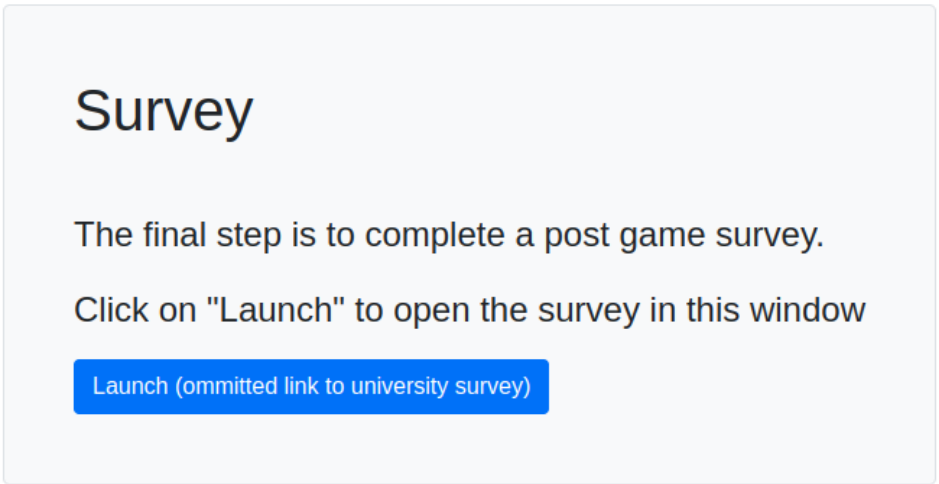
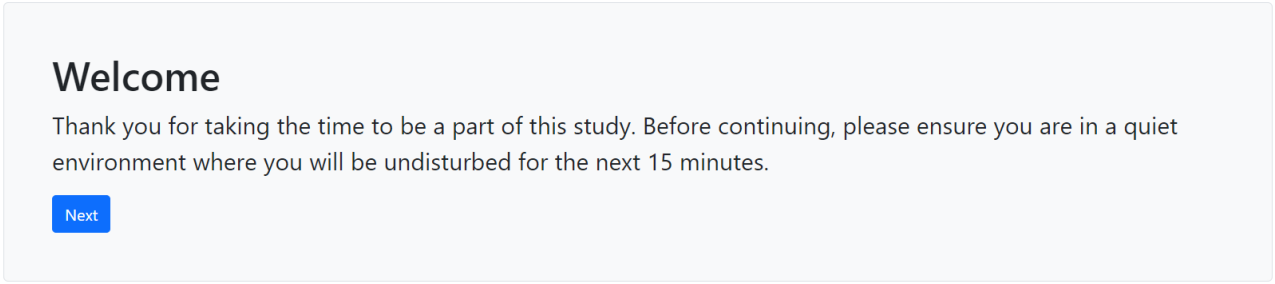


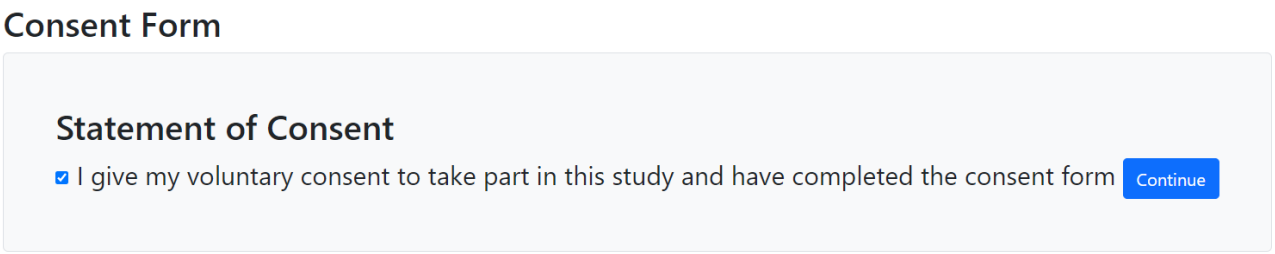
Figure 9: The teacher is able to see all game objects and use their speech to provide feedback for both the trial and actual games. However, they are unable to control the agent

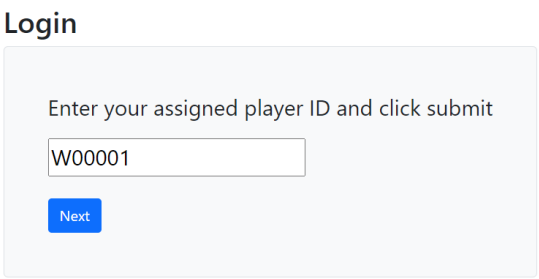


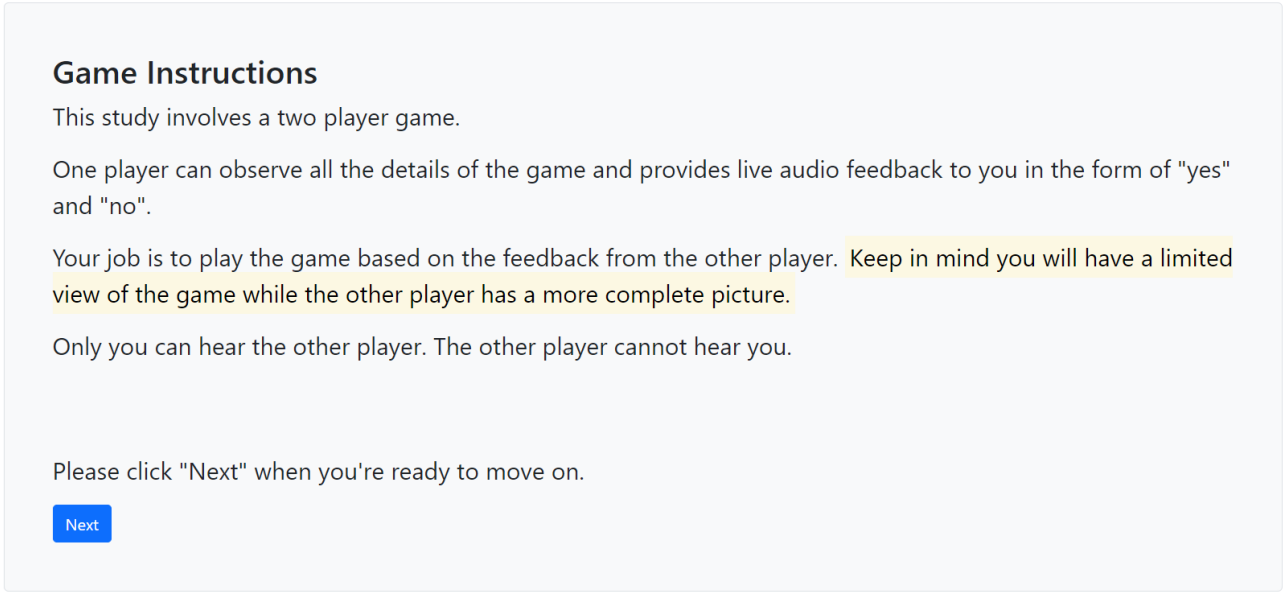
A.2 Data Collection interface (Wizard)

Below we share screenshots of the data collection interface provided to the wizard during a user study session.









Instructions

You will use the arrow keys (UP, DOWN, LEFT, RIGHT) to select an action.

If you do not select an action in a time step, the agent will take random actions, so **please continue striking keys to keep moving based on your beliefs and feedback from the other player**

Also, please wait until the robot has made a step after pressing a key before continuing.

Round 0 will be a practice round. You will be able to see important locations on the map only during the practice. After that, you will get a new map. Rounds 1-3 will use the same map but you will start in a different location.

Please click "Next" when you're ready to move on.

Next

Instructions

Key Points

- Your job is to play the game based on feedback from the other player. Try to complete each round as quickly as possible.
- Use the arrow keys to play the game.
- **After the practice round, the map will be randomized once. Rounds 1-3 will use the same map.**
- Try to figure out and remember the positions of important locations on the map. We will ask about these later.
- Do not use the back button in the browser. Let the study coordinator know if anything goes wrong.

Please click "Start Practice" when you're ready to start the game.

Start Practice

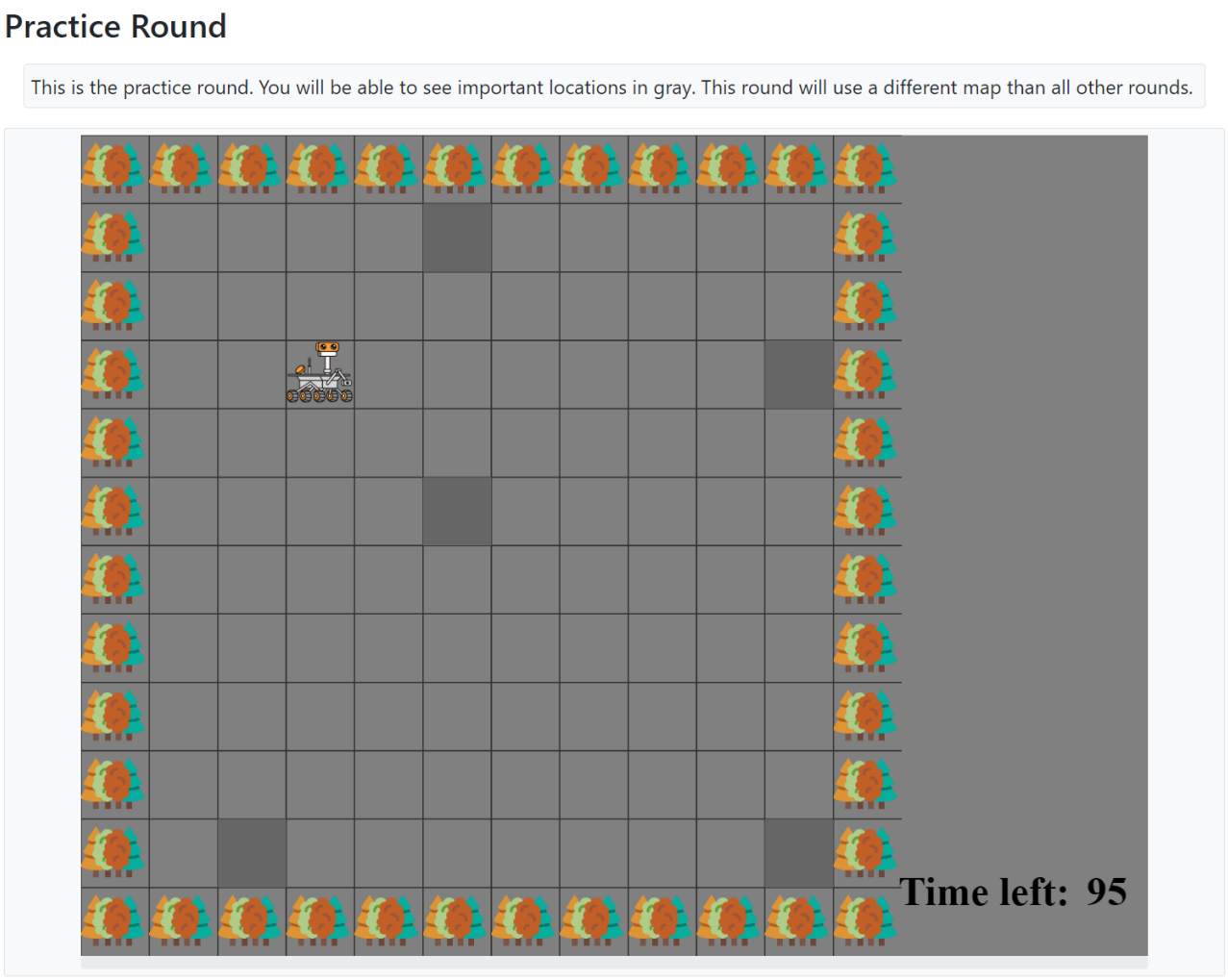
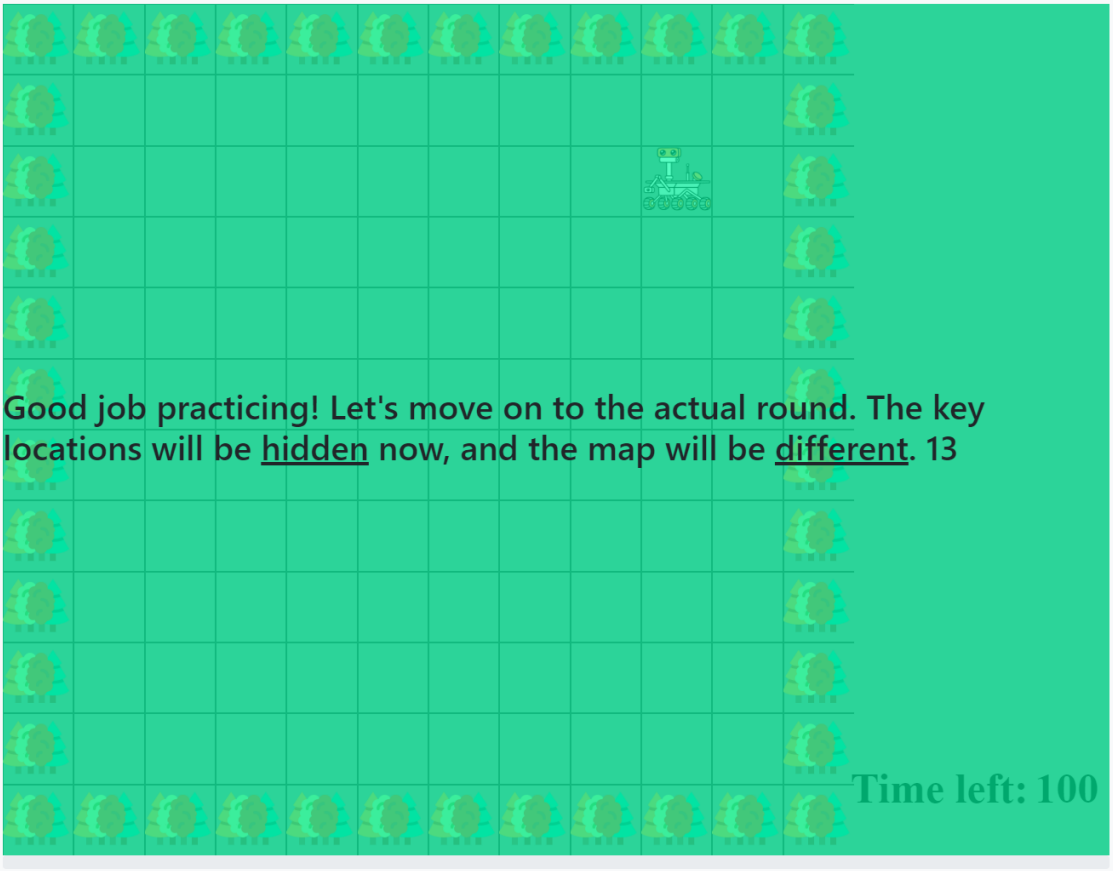


Figure 10: In the trial period, the worker is only permitted to see the locations of the game objects but not the objects themselves

Round 1 of 3

This round will use a DIFFERENT map. The important locations are hidden now. Try to figure out and remember their positions while playing the game.



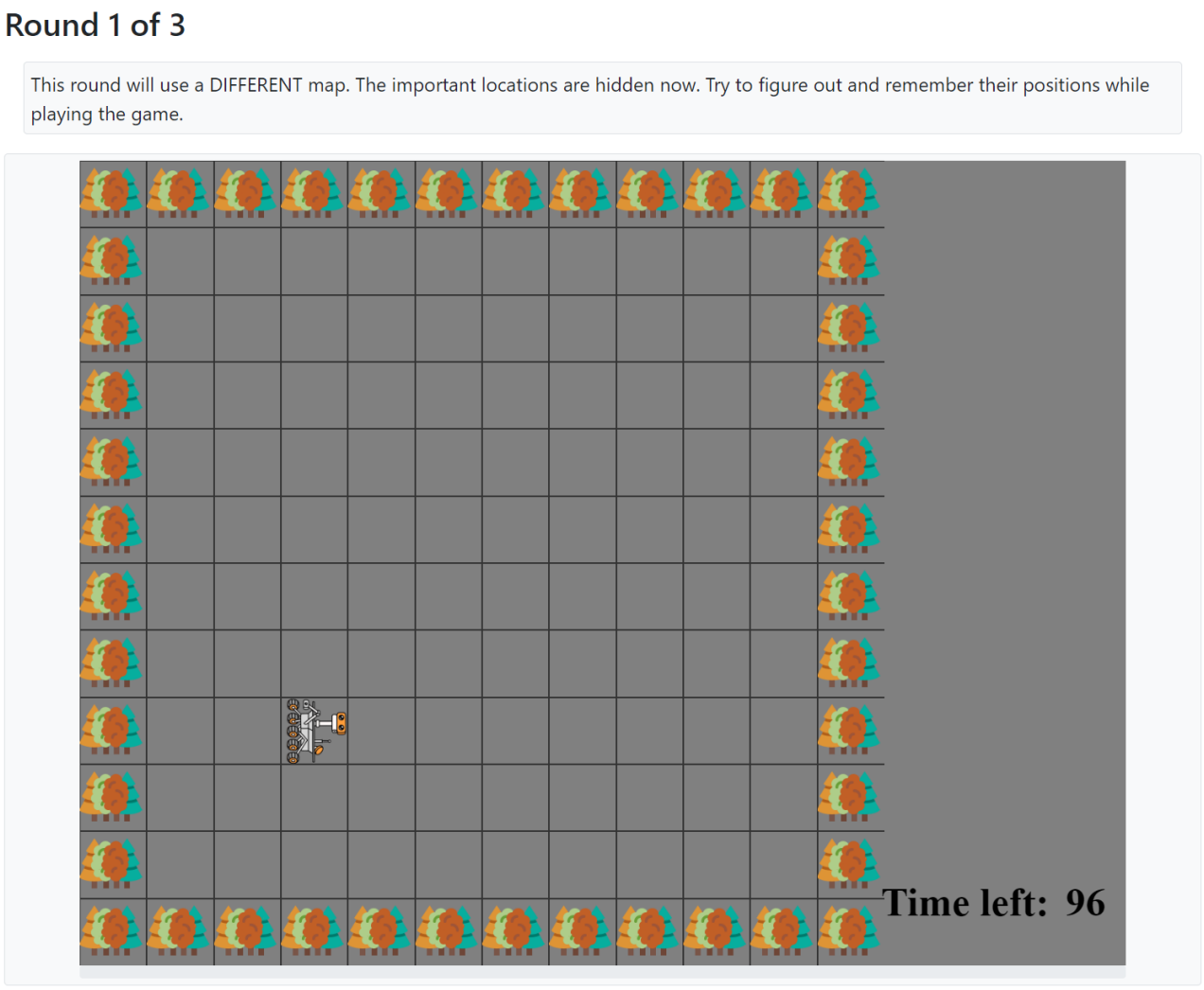
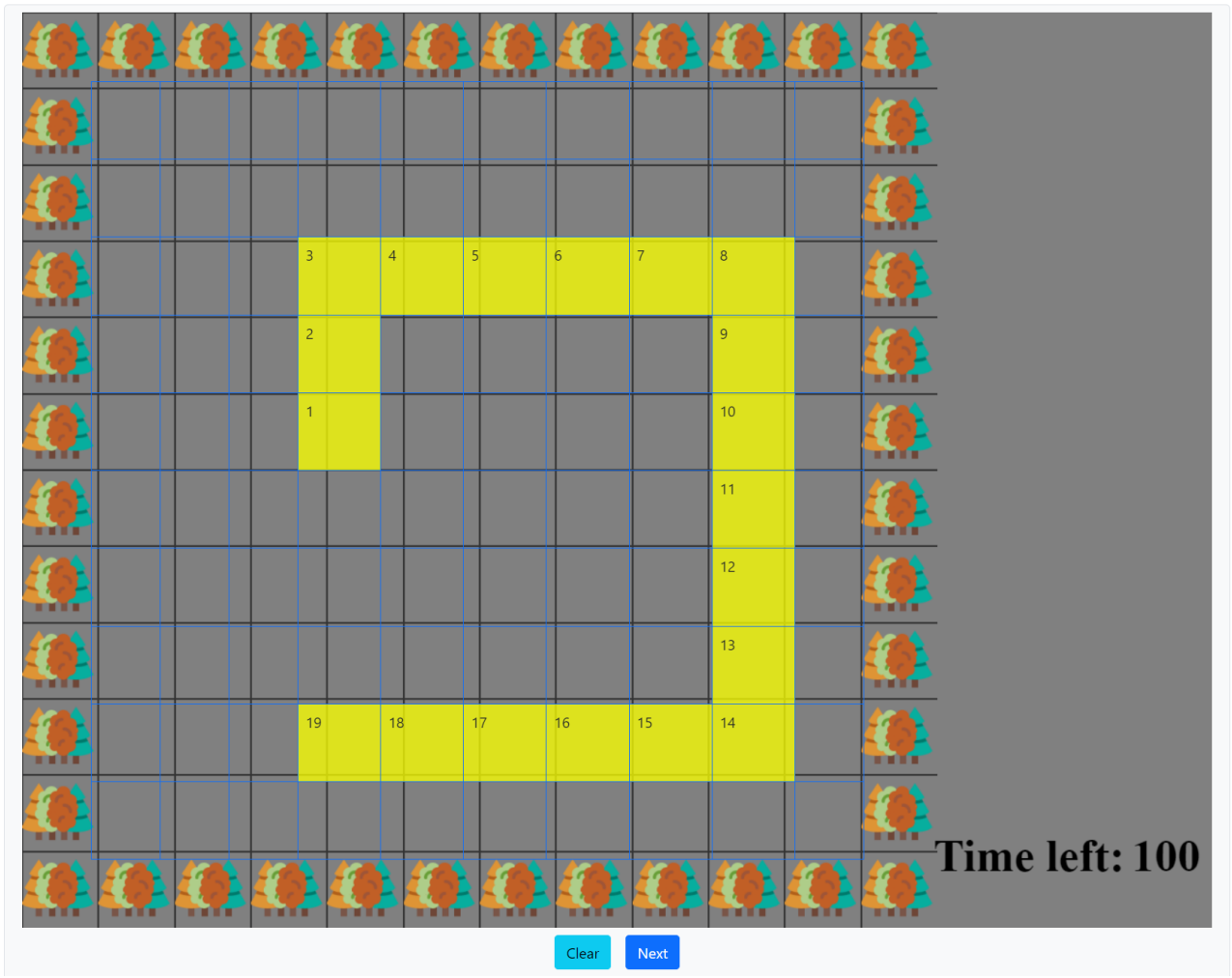
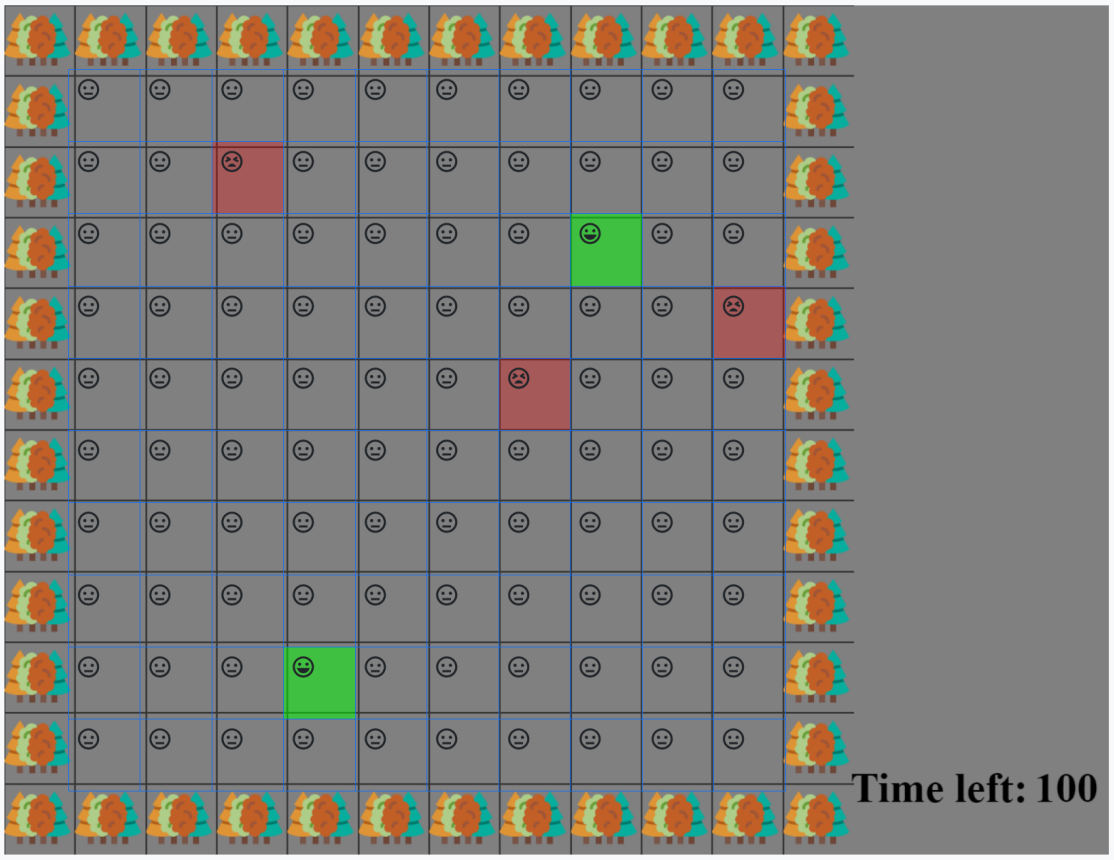


Figure 11: In the main recording session, the worker is unable to see any game objects nor their location


Round 3 of 3
Given that the rover starts in the yellow square, click on the sequence of squares that form the optimal path to the goal.You can go over the same square more than once. To the clear path, click "Clear". When finished, click "Next"
3
4
5
6
7
8
2
9
1
10
11
12
13
19
18
17
16
15
14
Time left: 100
Clear
Next

Round 3 of 3

Please mark the squares as good, neutral, or bad. Note that a majority of squares may be neutral, with only a small number being good or bad. Click on a square to cycle between good, bad, and neutral



Survey

The final step is to complete a post game survey.

Click on "Launch" to open the survey in this window

Launch (ommitted link to university survey)

A.3 TAMER Implementation

The implementation followed the instructions by Knox and Stone [19]. This included a discount factor of $\gamma = 0$. Therefore, the agent was completely myopic and only learned about actions that produce an immediate reward [19]. Additionally, the MDP reward function R was replaced by a human reward function H, which was completely based on the feedback given by the human trainer. The human reward function was learned by a supervised learning model [19]. A kernelized regression model was implemented with a kernel union to cover different gammas and variances. The library scikit-learn was used for the implementation of the regression models and the kernel featurizer. Moreover, the weights of the regression model were updated with a stochastic gradient descent algorithm and a learning rate of 0.01.

Following the approach of Deep-Q learning [45], the output of the human feedback model represents state-action values. We implemented four supervised models, one for each action, since regression models only give one output scalar. Each human feedback model took a vector representing the agent's current state as input. The state was the row and column of the agent's location as well as a bool whether the nut had been collected or not. The output value represented the state-action value for the input state, and the action represented by the supervised model. The agent learned H by updating one specific action model once a human reward was given for one state-action pair. If no reward was obtained and $h = 0$, the model(s) were not updated.

Furthermore, we implemented credit assignment as described by Knox and Stone [19]. Credit assignment splits the reward over the n most recent state-action pairs. As suggested by the authors, we implemented a probability density function (pdf) with gamma(2.0, 0.28). With the help of the pdf, weights for each of the state-action pairs are calculated, which represent the probability that the reward was meant for that pair. The weights were calculated for each state-action pair $t_1, t_2, ..., t_n$, where if $i < j$, then i happened before j. Thus, t_n was the most recent timestamp. The timestamp of when the human feedback was obtained was denoted with 0 on the gamma pdf. Then, the weights were iteratively calculated for each state-action pair by taking the integral $\int_{t_{i-1}}^{t_i} f(x)dx$, where t_i represents the timestamp when the agent executed the action. The resulting weight, hence the probability that the reward was meant for that specific state-action pair, is then multiplied by the obtained human reward and used to update the human reward function H. We chose $n = 3$ because of the relatively low speed of the robot. Therefore, it was quite unlikely that the feedback was meant for an older agent timestamp.

Since the agent learned with offline learning, it did not choose its own actions. Instead, H was updated based on the recorded state-action pairs and voice feedback of the WoZ experiment.

A.4 System Design for WoZ data collection pipeline

We show a schematic diagram (Fig. 12) of the system design for the WoZ data collection setup described in Sec. 3.1.

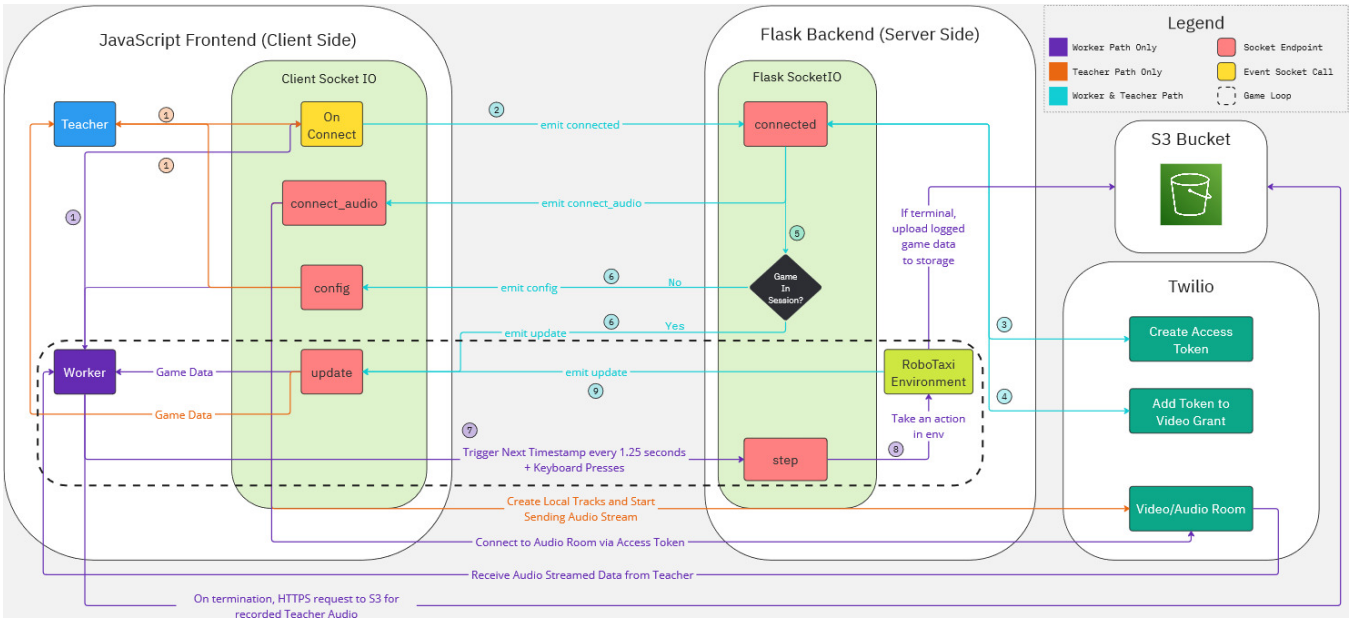


Figure 12: System Design for the Wizard of Oz data collection setup for the interactive RL setup.

B APPENDIX: AUDIO-AUGMENTED DEMONSTRATIONS

B.1 Additional Results

We show additional results from the analysis of the user study data for three different Atari games below (Fig. 13-16).

B.2 Screenshots of Atari

Below we share screenshots of the data collection interface provided to the demonstrator during a user study session for one of the three games (Fig. 17-18).

Received 10 May 2024

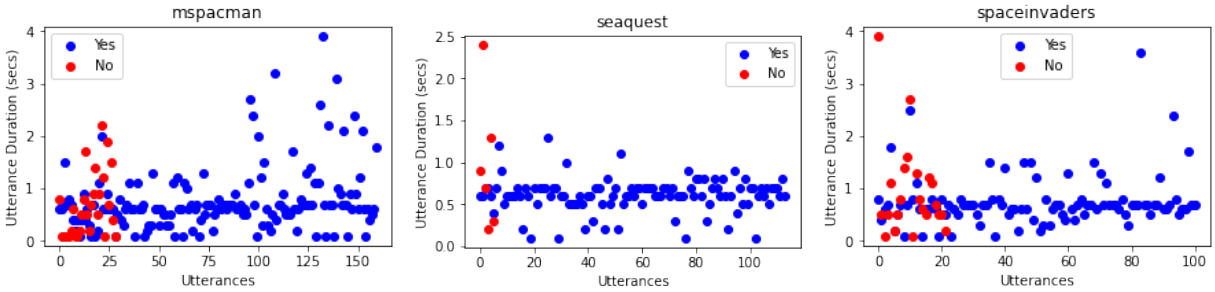


Figure 13: Duration (in seconds) of yes and no utterances from a single demonstrator, during demonstrations provided to three Atari games.

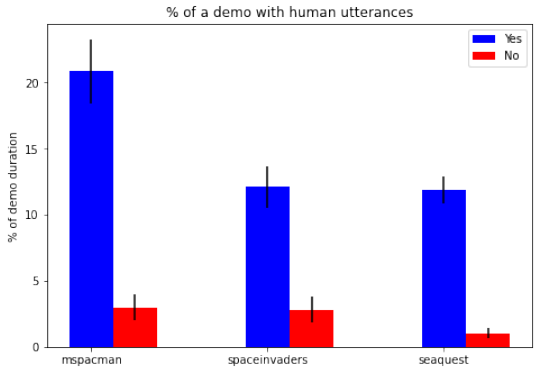


Figure 14: Percentage of yes and no utterance duration given the total demonstration duration.

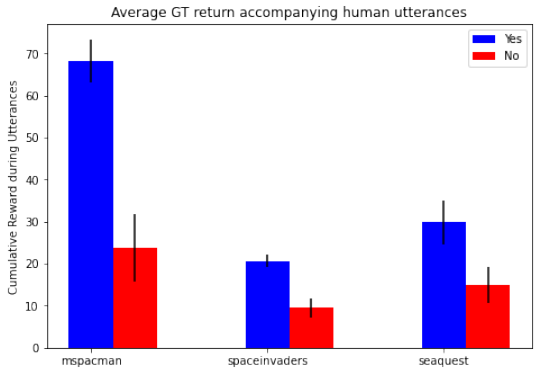


Figure 15: Cumulative reward of snippets surrounding audio utterances.

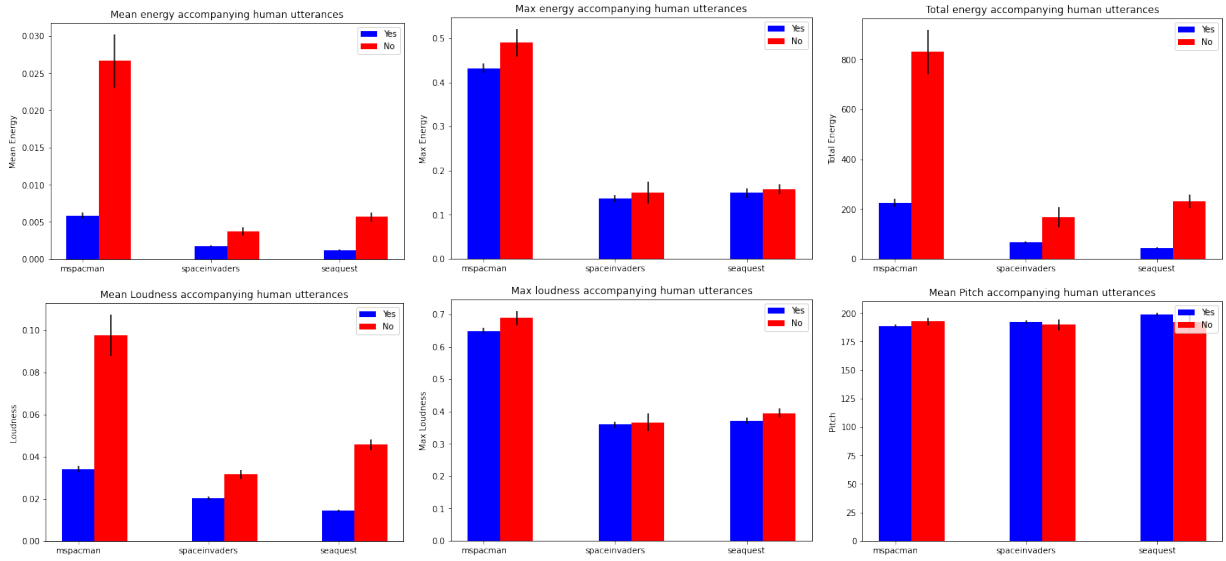


Figure 16: Prosodic feature values for yes and no utterances from a single demonstrator providing demonstrations to three different Atari games.

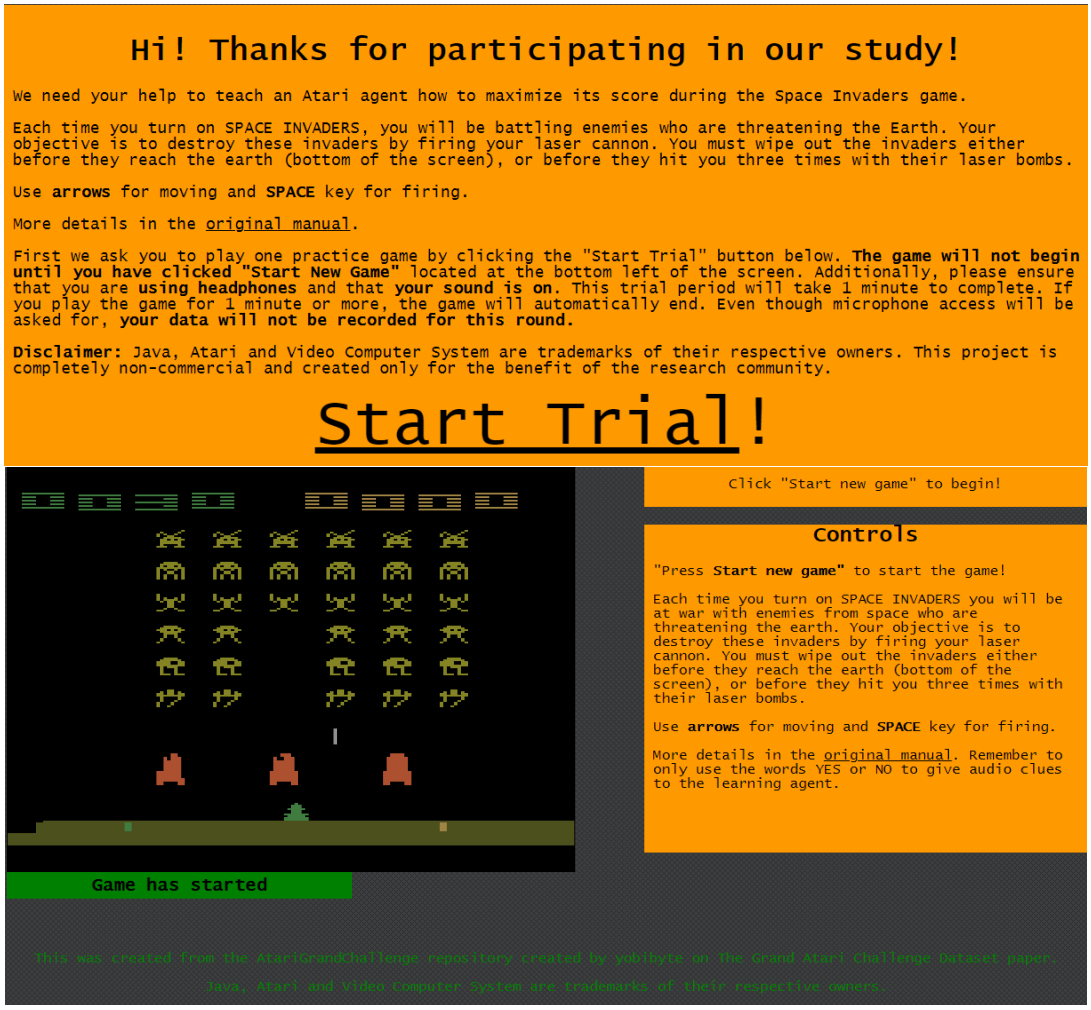


Figure 17: Trial period for the participant to get accustomed to the controls of the selected Atari game

Figure 18: Game period to record the participant's voice and game data as they play an Atari game