NatSGD: A Dataset with Speech, Gestures, and Demonstrations for Robot Learning in Natural Human-Robot Interaction

Snehesh Shrestha, Yantian Zha, Ge Gao, Cornelia Fermuller, Yiannis Aloimonos

University of Maryland College Park
snehesh@umd.edu

Abstract

Recent advances in multimodal Human-Robot Interaction (HRI) datasets feature speech and gesture. These datasets provide robots with more opportunities for learning both explicit and tacit HRI knowledge. However, these speech-gesture HRI datasets focus on simpler tasks like object pointing and pushing, which do not scale easily to complex domains. These datasets focus more on collecting human command data but less on corresponding robot behavior (response) data. In this work, we introduce NatSGD, a multimodal HRI dataset that contains human commands as speech and gestures, along with robot behavior in the form of synchronized demonstrated robot trajectories. Our data enable HRI with Imitation Learning so that robots can learn to work with humans in challenging, real-life domains such as performing complex tasks in the kitchen. We propose to train benchmark tasks that enable smooth human-robot interaction, including 1) language-and-gesture-based instruction following and 2) task understanding (Linear Temporal Logic formulae prediction) to demonstrate the utility of our dataset. The dataset and code are available at http://snehesh.com/natsgd website.[1]

Introduction

Humans communicate with each other by spontaneously using language and gestures (*p*.230 in (Tomasello 2010)). The reason is that language-based information is effective at communicating explicit knowledge, while more tacit knowledge and behavior-based communication is effective at conveying tacit knowledge but often too costly to represent explicit tasks. Moreover, many everyday tasks, such as cooking and cleaning, are a mixture of explicit and implicit information.

If robots could collaborate in a human-like (i.e. more natural) way, it would reduce the cognitive load for humans. This points to a need for datasets that can help robots learn to effectively incorporate more natural human advice into both language and gestures. However, the majority of HRI datasets that enable fantastic collaborative robots like Google Home, Amazon Alexa, and Apple Siri, solely rely on speech as a communication channel (Novoa et al. 2017; James, Tian, and Watson 2018; Vasudevan, Dai, and Van Gool 2018; Narayan-Chen, Jayannavar, and Hockenmaier 2019; Padmakumar et al. 2022). On the other hand, many other HRI datasets like (Pisharady and Saerbeck 2015; Shukla, Erkent, and Piater 2016; Mazhar et al. 2018; Chen et al. 2018; Chang, Tejero-de Pablos, and Harada 2019; Gomez Chavez et al. 2019; Neto et al. 2019; Nuzzi et al. 2021; de Wit, Krahmer, and Vogt 2021), only focus on gestures when humans interact with robots (Luan et al. 2016). Some works propose HRI datasets involving both speech and gestures (Matuszek et al. 2014; Rodomagoulakis et al. 2016; Azagra et al. 2017; Chen, Leu, and Yin 2022). However, these works focus on the perception aspect of HRI tasks like recognizing which object a human is referring to in the form of simple colors and shapes and tasks like picking and placing. For complex real-world objects and tasks, people use a wide variations of features, vocabulary, and styles to refer to objects and actions.

Therefore, in this work, we aim at creating a novel HRI dataset that: 1) resembles natural communications including both speech and gestures, 2) helps robots learn complex tasks like cooking and cleaning that are valuable in our everyday life, and 3) includes demonstration trajectories due to the complexity of tasks. Specifically, we designed a wizard of Oz (WoZ) experiment (Dahlbäck, Jönsson, and Ahrenberg 1993) where participants interacted naturally with what they believed was a highly capable autonomous humanoid. We use the outcome of these experiments to form our main contribution, NatSGD, a natural communication dataset (Fig. 1) that provides novel missing speech and gesture data along with paired robot demonstration towards task completion. Furthermore, NatSGD consists of data from the point of view of the speaker, states of the objects and the robot, human-annotated speech, gesture, gesture types, the body parts involved, the corresponding robot trajectories generated from expert teleoperations, robot's point-of-view (Ego View), depth, scene and object instance segmentation, and semantic segmentation (see Fig. 2). *NatSGD* has 1143 commands given by 18 people with 11 actions, 20 objects, and 16 states. We propose two important HRI benchmark tasks, Language-and-Gesture-Based Instruction Following and Human Task Understanding, to demonstrate a great potential of promoting fundamental HRI research with our dataset.

To the best of our knowledge, our *NatSGD* dataset is the

[1] http://snehesh.com/NatSGD

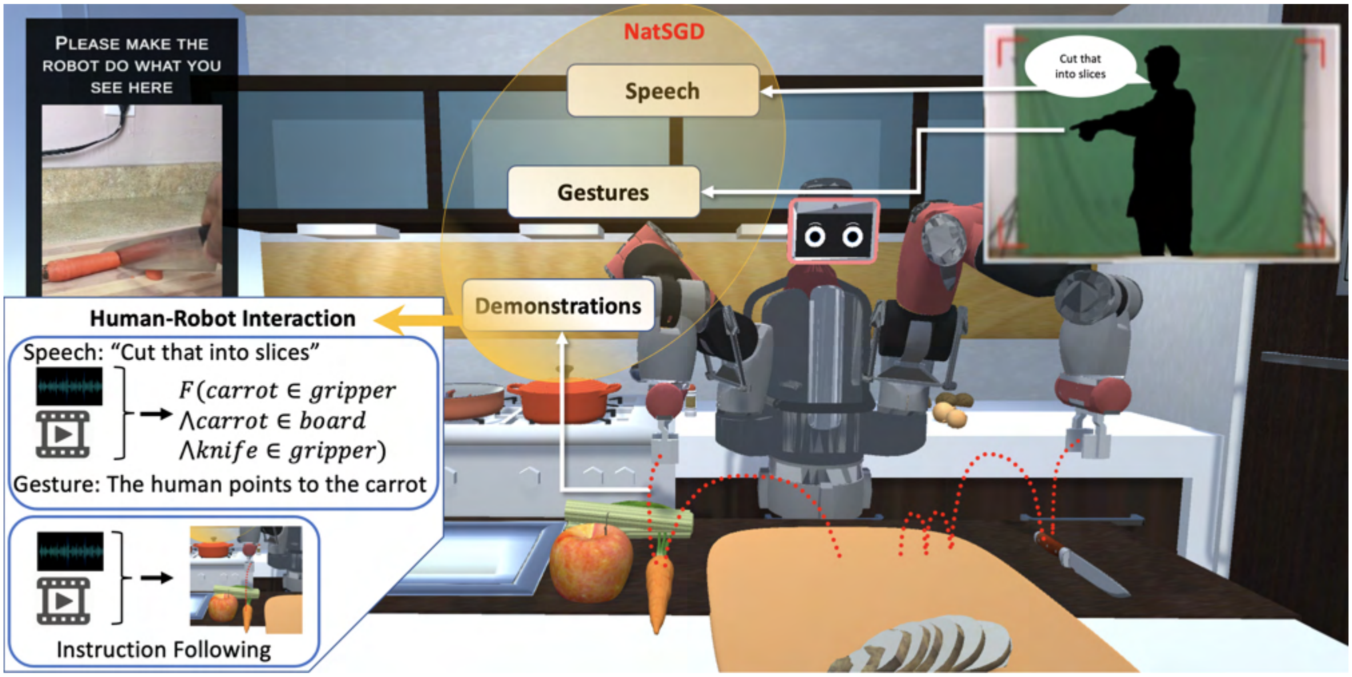

PLEASE MAKE THE ROBOT DO WHAT YOU SEE HERE
NatSGD
Speech
Gestures
Demonstrations
Cut that into slices
Human-Robot Interaction
Speech: "Cut that into slices"
$F(carrot \in gripper$
$\wedge carrot \in board$
$\wedge knife \in gripper)$
Gesture: The human points to the carrot
Instruction Following

Figure 1: NatSGD contains speech, gestures, and demonstration trajectories for everyday food preparation, cooking, and cleaning tasks. The NatSGD dataset potentially enables the learning of complex human-robot interaction tasks due to the rich interaction modalities and strong supervising signals at both trajectory-level (demonstrations) and symbolic-level (ground-truth activities that match humans' intention).

first HRI dataset that includes speech, gestures, and demonstration trajectories for robots to learn complex tasks during cooking and cleaning – tasks that are ubiquitous in our everyday lives.

The paper is structured as follows: Section 2 (Related Work) will compare and contrast other HRI datasets and highlight the need for our dataset. Section 3 (NatSGD Dataset) will provide a detailed account of the design and the careful considerations in the creation and annotations of the dataset. Section 4 (Evaluation Tasks) will include the benchmarks tasks that are proposed. In section 5 (Conclusion), we summarize our work and highlight our contribution.

Related Work

As mentioned in the Introduction section, there are various datasets that adopt either speech or gestures as communication modalities. In this section, we focus on discussing the datasets that involve both communication modalities. We further discuss how those datasets can enable the learning of HRI systems.

Datasets for Speech-and-gesture-based HRI Many Speech-and-Gesture datasets do not involve a robot. For example, the HRI datasets presented in (Azagra et al. 2017) and (Rodomagoulakis et al. 2016) completely focus on perceptual tasks. (Azagra et al. 2017) aims at solving tasks like interaction recognition, target object detection, and object model learning, whereas (Rodomagoulakis et al. 2016) addresses tasks like visual gestures recognition and audio command phrases recognition. (Lee et al. 2019) and (Kucherenko et al. 2021a) presents gesture-speech multimodal human-human conversation datasets that have a potential of training robots for HRI scenarios. Unlike the above datasets, there are also a few that demonstrate values for robotics research. (Matuszek et al. 2014) introduces a HRI dataset in which humans use both gesture and language to refer to an object. Robots learn to recognize the referred object and push that object away. Similarly, (Chen, Leu, and Yin 2022) collects a speech and gesture dataset and demonstrates its value on a robotic pick-and-place task in the HRI settings. The pick-and-place system works by having robots recognize target locations by observing humans' speeches and gestures and then inputting the target region to a motion planner. In contrast to these datasets, our NatSGD dataset includes demonstration trajectories and thus enables the robot learning of more complex tasks like cooking and cleaning. The importance of including demonstration trajectories in HRI datasets is also acknowledged by a recent dataset paper (Padmakumar et al. 2022), which however does not consider gestures.

Robot Applications on HRI Datasets The HRI community also lacks robot learning works that involve training deep neural networks on HRI datasets. The aforementioned HRI works (Matuszek et al. 2014; Chen, Leu, and Yin 2022) put more focus on the perceptual aspects of using their collected datasets and therefore only applies perception outputs (e.g. the pose of recognized object) to an engineered robotic system. Similarly, (Krishnaswamy and Pustejovsky 2020) provides a formal analysis on the multimodal recognition of referred objects in HRI. (Ahuja et al. 2020; Habibie et al. 2021; Kucherenko et al. 2021b; Habibie et al. 2022) research the correlation between speech and gestures that could lead to better gesture recognition or generation functions. However, these works do not really involve robots. (Codevilla et al. 2018) introduces an imitation learning framework that is conditioned on human language commands. Their framework allows the behavior of autonomous cars changed by human users in testing phase. In contrast to this work, we evaluate our dataset on imitation learning that is conditioned on both language and gesture commands for instruction following.

NatSGD Dataset

NatSGD Dataset aims to address gaps in the current datasets to address naturalness and usefulness from research to real life application. For such a human-robot interaction dataset for robot learning to be useful, it has to be complete and comprehensive. Ideally, on the human side, the participants and their behavior needs to be controlled for bias. On the robot side, the robot needs to be capable and the simulator needs to be *realistic* so that the humans interacting with it will believe and trust it. Interactions should be *natural* so that the behavior spontaneously emerges based on subject's own free will, without being primed by the researchers. And applicability of the data needs to be *versatile* such that it can be used for many robot learning tasks. To this end, the following sections highlight the key considerations and aspects of NatSGD that make it novel and useful.

Humans and Bias

It is important for the data to be fair and well understood. People have implicit and explicit biases (Banaji and Greenwald 2013) and data can inherit these biases. It is important to control for biases from both technical machine learning perspective, as well as, the societal implication perspective. Human background such as gender, age, expertise, culture, and personality were considered. Additionally, individual experience in the form of NASA-TLX work load (Hart 2006) and their impression of the robot were also recorded. Robot's gender, approachability, and naturalness were considered in the design of the robot face, name, and movements. These are detailed in the appendix Robot Face and Name section.

Eighteen subjects of ages 18 to 31 years (Mean 20.91±3.75), participated in this experiment (9 male and 9 female.) Their personalities were identified as Extroversion (5.56±2.09), Agreeableness (9.17±1.15), Conscientiousness (8.17±1.38), Emotional Stability (7.56±1.79), and Openness to Experience (7.89±1.47). Equal distribution of technical to non-technical background participants were chosen based on their first-hand exposure to high-tech games and toys and their education or profession as computer science and engineering. No participants had prior interaction with robots before the study. Please see appendix for more details in Participants section.

We transcribed human speech into text and extracted their word embedding using Glove model (Pennington, Socher, and Manning 2014). Similarly, we extracted human poses by using OpenPose (James, Tian, and Watson 2018) and used a sequence of human pose vector as the gesture low level feature.

Robot and Realism

Anyone who has worked with a real robot to accomplish a real life household task understands how difficult it is. Kitchen tasks like cutting a vegetable with a knife might be simple for humans, but they are extremely challenging for a robot. So, there are many challenges of conducting HRI experiments on a real robot. It is even more challenging when there is a real-time requirement such as collecting a natural interaction dataset. Inspired by recent progress in sim2real research (Abeyruwan et al. 2022; Kaspar, Osorio, and Bock 2020), we designed and developed our photorealistic simulator. We created a real-time robotics simulator using Unity3D (Unity Technologies 2021), a game engine, that can be be controlled by a researcher quickly based on the subject's commands. Therefore, HRI data with a simulator opens many opportunities for pushing forward HRI research especially when we have complex real-world environments and tasks. Our research also opens a door for future sim2real research that handles HRI settings.

The Simulator As shown in Fig. 2, the NatSGD Robot Simulator was built using Unity 3D along with modified ROS plugin (Bischoff 2021). The system ran on a computer with Intel i7 Gen 16GB RAM connected to the 55" TV in order to ensure smooth *realistic rendering* and processing in *real-time*. On the top right corner of the screen, a camera feed of the participant was overlayed as shown in Fig. 1. This served as a feedback mechanism to the participants to feel the robot could see them and to help them stay within the frame. The robot real-time inverse kinematics (IK) of the head movement and robot arms were implemented with BioIK (Starke et al. 2017; Starke, Hendrich, and Zhang 2018). The robot looked at the target object(s) to demonstrate robot's attention while performing a task. When the robot was ready to interact with the subject for the next command, the robot looked back at the participant. The detailed output from the simulator includes the human and robot perspectives. It also includes the object states and robot trajectories which are detailed in Ground Truth Labels section. Activities like pouring and cutting have causal sub-activities. For example, for cutting a tomato, the robot needs to locate a knife and grasp it. It then needs to approach the tomato on the chopping board, hold it stable, and finally make its first cut, then follow up with subsequent slices. These are seamlessly implemented in the simulator.

Interaction and Naturality

Data of natural interaction is important to capture human behavior. The robot can learn from these natural cues that are unstructured, mixed-modal, and consists of implied contexts. They even contains contradictory phrases and repairing mechanisms. To incite natural human behavior in the lab is challenging. We took inspiration from prior literature, conducted a number of pilot studies, and came up with WoZ

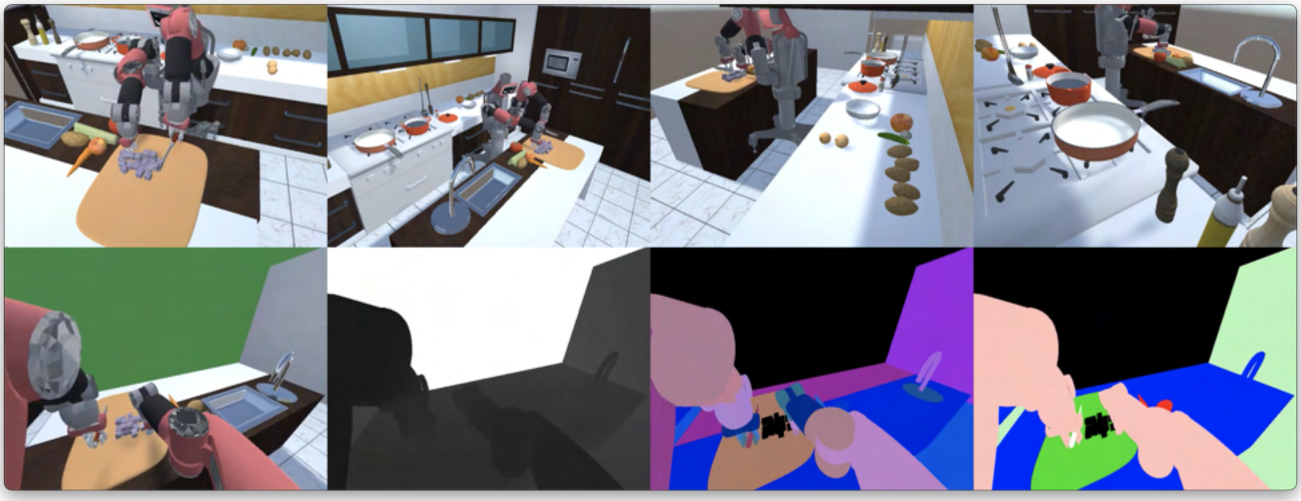


Figure 2: Our Simulator that shows Baxter cutting onions into pieces. The simulator includes multi-view perspectives of the scene and the robot from fixed and moving cameras. The top row shows human-first-person view, the ceiling top left, the counter bottom right, and counter bottom left cameras in the kitchen. The bottom row shows robot's egocentric view in RGB, depth, unique object segmentation, and category based semantic segmentation. These perspective are useful for robot to learn to complete the tasks based on human speech-gesture commands and it's own observation of the world and object states.

experiment design to incite natural emergent human behavior.

Pilot Studies To achieve the best of both worlds of in-the-wild and controlled lab study, our WoZ experiment design deceives the participants into believing the robot is fully autonomous. We then conducted multiple pilot studies to validate factors that could affect participant behavior to validate independent and dependent control variables as well as the workflow. We experimented with the (a) background noise (see appendix Background Noise section), (b) perceived robot personality and capability based on the robot's face and name (see appendix Robot Face and Name section), (c) staging to keep the participants engaged (see Staging section), (d) considerations of the priming effects from practice sessions (see appendix Practice Session section), (e) WoZ clues that participants might be able to use to figure out the hidden agenda (see appendix WoZ Cues section), and (f) the effect of experiment instructions (see appendix Instruction section). These findings informed our experiment design decisions.

Expert Demonstration and WoZ Control Policy For both practice and the data collection sessions, the researcher facilitator (wizard) controls the workflow to move ahead as long as the participant's command is related to the task at hand and is discernible. Robot nods "yes" (head up and down motion) on commands that robots understood. For unintelligible or unrelated commands, the robot displays the confusion face. For example, if a participant gestures to move right, or say move right, the wizard makes the robot move right. However, if the participant mumbles and the wizard cannot hear the participant, the wizard prompts with a confused robot face for clarification from the participant. The clarification repair prompt (e.g. a confused face animation) is intentionally designed to be ambiguous so that the participant attempts different strategies in providing instructions for the same task. The participant is allowed a maximum of five attempts per command. If unsuccessful after that, the task is skipped. In rare cases, the task can also be skipped, if the Unity game becomes frozen or Baxter is not able to solve the IK for grasping objects for more than 30 seconds.

Staging As the participant and the robot do the share the same immersive space, one challenge that we observed was when the robot was not directly facing the person. The participant sometimes loses the context, visibility, and frame of reference. To account for this, we borrow techniques from the 12 principle of animation, specifically *staging* (Thomas, Johnston, and Thomas 1995) to gently move the camera to a camera pose that gives a clear view of the key event. For example, if the robot is pouring oil into the pan, we pick camera angle 2 such that the pan is in the center, with the robot in the background, and oil visible to one side of the screen.

Versatile Applicability

We believe the dataset would be more useful if it can be applied towards multiple use cases. To this end, NatSGD consists of data in multiple modalities and multiple vantage points from multiple subjects. It consists of long continuous sequences with ground truth labels that are computationally generated or annotated by multiple human annotators.

Supported Learning Tasks NatSGD can be used for low-level tasks such as gesture recognition, speech recognition, and object detection. For gesture recognition, a semantic level recognition e.g. pointing to an onion is extremely useful when paired with speech ”cut it into pieces.” Additionally, gesture property recognition (intentional or unintentional) classification is a practical applications in HRI. An unintentional gesture could be a subject stretching which could be mistaken for pointing. More importantly, NatSGD can be used for high-level robot learning tasks such as task understanding and instruction following. The high level tasks are proposed to be evaluated as benchmark tasks for this dataset and is elaborated in the Evaluation Tasks section.

Ground Truth Labels The structure of this dataset has been from a utilitarian perspective for robotics and machine learning applications. NatSGD contains labels for each task that were completed, for eg. cutting an onion or a tomato. For each task, we also break them down into labels of the their sub-task such as grabbing the knife, holding the onion, and cutting are provided. We include synchronized data from both robot and human perspectives.

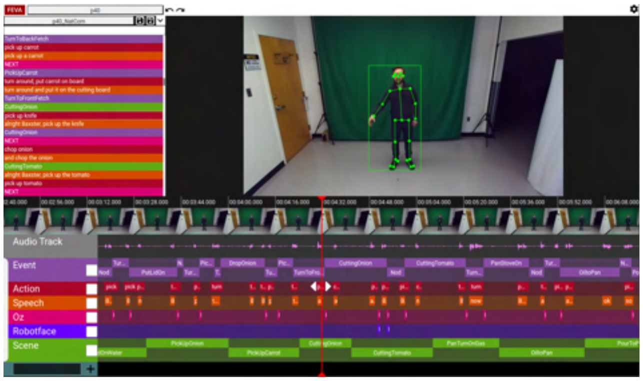
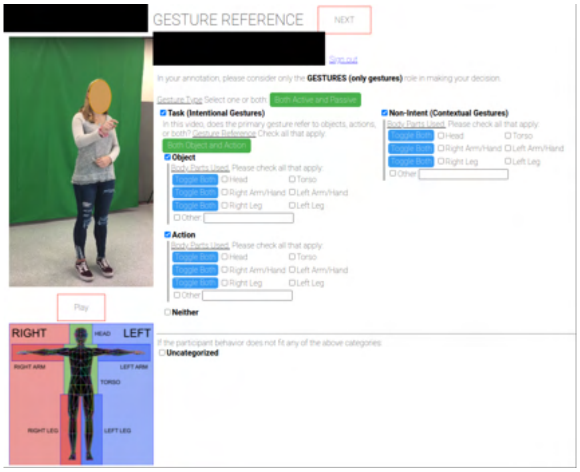

FEVA
p40
p40_NatCom
TurnToBackFetch
pick up carrot
pick up a carrot
NEXT
PickUpCarrot
turn around, put carrot on board
turn around and put it on the cutting board
TurnToFrontFetch
CuttingOnion
pick up knife
alright Baxter, pick up the knife
CuttingOnion
NEXT
chop onion
and chop the onion
CuttingTomato
alright Baxter, pick up the tomato
pick up tomato
NEXT

Figure 3: Example of temporal annotation of the tasks, subtasks, speech, and gesture of participant p40 with Fast Event Video Annotation (FEVA (Shrestha et al. 2023)) tool.

On the human side, we temporally segmented human commands as shown in Fig. 3. Based on *modality*, each command was annotated for consisting relevant *speech* and *gesture*. We then further distinguished them based if they *referred* to *objects* or *action*. The gesture is annotated as containing task-related (*intentional*) and task-unrelated (*unintentional*) gestures. Finally, for all gestures, we also annotated the role of each *body part* in task specific action/ object reference and non-intent movements as illustrated in Fig. 5. Please see appendix for more details in the Ground Truth Labels section.

On the robot side, sequences of images are temporally synchronized across all cameras along with the audio. The depth images are generated by using the normalized distance from the robot's egocentric view. Additionally, individual object's unique segmentation and their object semantic segmentation, based on their object categories such as food, utensils, and appliances are also provided. NatSGD also includes expert trajectories of the end-effectors, head, and the robot base that demonstrates how to perform the tasks, head control, and robot navigation.


[redacted]
GESTURE REFERENCE
[redacted]
Sign out

In your annotation, please consider only the **GESTURES (only gestures)** role in making your decision.

Gesture Type Select one or both: Both Active and Passive

☑ **Task (Intentional Gestures)**
In this video, does the primary gesture refer to objects, actions, or both? Gesture Reference Check all that apply:
Both Object and Action
☑ **Object**
Body Parts Used. Please check all that apply:
Toggle Both ☐ Head ☐ Torso
Toggle Both ☐ Right Arm/Hand ☐ Left Arm/Hand
Toggle Both ☐ Right Leg ☐ Left Leg
☐ Other
☑ **Action**
Body Parts Used. Please check all that apply:
Toggle Both ☐ Head ☐ Torso
Toggle Both ☐ Right Arm/Hand ☐ Left Arm/Hand
Toggle Both ☐ Right Leg ☐ Left Leg
☐ Other
☐ **Neither**

☑ **Non-Intent (Contextual Gestures)**
Body Parts Used. Please check all that apply:
Toggle Both ☐ Head ☐ Torso
Toggle Both ☐ Right Arm/Hand ☐ Left Arm/Hand
Toggle Both ☐ Right Leg ☐ Left Leg
☐ Other

If the participant behavior does not fit any of the above categories:
☐ **Uncategorized**

Figure 4: Gesture Type, Reference, and Body Parts Annotation of participant p61 using FEVA Crowd tool.

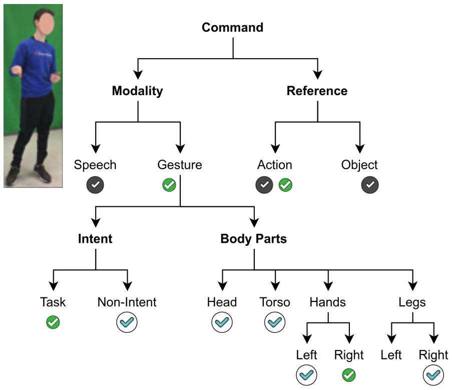


Figure 5: In the example above, participant p58 says "Baxter, stir the pot", with right hand showing the stirring gesture while the body shifts to his left with head and left hand moving that could be read as being unsure. Here 'stir' is speech action, 'pot' is speech object, while right hand is gesturing for stirring task action, the head, torso, left hand and right leg are unintentional gestures.

Reliable and Robust Data We carefully take iterative steps to prepare the data to ensure integrity, quality, validity, and fairness. We *calibrated* and *synchronized* all the cameras audio-video. We *compressed* the final video to a lossless high quality data format. Multiple people *annotated* our data using Fast Event Video Annotation Tool (FEVA (Shrestha et al. 2023)) (see Fig. 3) and FEVA crowd (see Fig. 4). We annotated each label at least twice with up to three rounds of agreement checks for purposes of speed and reliability. We computed the *inter-rate reliability (IRR)* (Cohen 1960) and our annotation was updated in three rounds. Detailed breakdown of the IIR and the statistics are available in the appendix in the Inter-rater Reliability section.

Evaluation Tasks (Proposed)

We propose to evaluate our dataset on several benchmark tasks that are fundamentally important to enable smooth human-robot interaction. In particular, we consider 1) Robot learns to follow humans' instructions in the forms of speech and gestures; and 2) Understanding the task that is specified by a pair of human speech and gestures – formulated as learning a mapping from a pair of speech and gestures to a Linear Temporal Logic formula.

Language-and-Gesture-based Instruction Following

To the best of our knowledge, there is no robot instruction following works that takes both of speech and gestures at the same time. We therefore decide to adapt a work of language-conditioned imitation learning for robot manipulation (Stepputtis et al. 2020) to take gestures and speeches together. Their work uses semantic data from GloVe embedding (Pennington, Socher, and Manning 2014) for language, Faster RCNN (Ren et al. 2015) for images along with a single arm robot trajectories for control steps. Their work uses simple pick and pour tasks and simpler objects with two shapes, two sizes, and five colors with speech in the form of structured typed texts. To scale this to everyday real-life complex scenarios, NatSGD contains 11 distinct tasks and 20 objects with variety of features. NatSGD contains unstructured natural human speech utterances. It also consists of natural human gestures including non-communicative/ irrelevant gestures performed with relevant ones. Both speech and gestures appear as sequences of single or multiple commands. NatSGD contains two arms manipulation trajectories, head tilt trajectories, and the whole body navigation trajectories. In this benchmark task, we use robot ego view image, object segmentation, human speech, and human gestures to learn to generate location and manipulation trajectories to complete the commanded task.

Human Task Understanding

Because we collect our dataset in a way how communications can naturally happen as if humans talking to each other, there is no one-to-one mapping between a pair of a speech and a gesture to a single activity in our dataset. For example, when a participant says "pour soup into the bowl", it implies multiple conditional sub-tasks such as fetching the bowl and placing it close the pot. If the pot is covered, it needs to be uncovered. It then needs to find the ladle, scoop some soup, move the ladle without spilling over the bowl, and pour into the bowl. And this needs to be repeated until a desired amount is reached.

Therefore, we consider learning a mapping from speech and gestures to a Linear Temporal Logic (LTL) formula that describes a task (Pnueli 1977; Kesten, Pnueli, and Raviv 1998; Konur 2013). LTL is a modal logic system that can specify temporal relations among events (in the form of formulae) and do logic reasoning based on those formulae. Robotics and planning researchers have adopted LTL to formulate high-level reactive task specifications (Finucane, Jing, and Kress-Gazit 2010; Guo, Johansson, and Dimarogonas 2013; Baran et al. 2021). An example is that if we have a natural language task specification "*Go to all rooms on the first floor and then go to the second floor*", we can convert it into an LTL formula as "$F(\forall \text{room} \in \text{rooms } \mathsf{s.t.}\ \text{floor}_1 \bigwedge XF(\text{floor}_2))^1$", where F and X denote *finally* and *next* respectively. With an F operator, all variables that are inside F must hold True to make the formula satisfied. That said, the language-LTL conversion becomes non-trivial in natural human-robot interaction settings. To truly infer the task that the human specifies, a robot should also consider humans' body language and the robot's current state. Take the speech in Fig. 1, "Cut that into slices", as an example. The robot needs to use the human's gesture of pointing to the carrot to understand what is "that" in the speech. Also, since the robot is not holding the knife, the human, while did not explicitly mention the knife, implies that the robot should grasp the knife. Therefore, the converted LTL formula in that scenario is: X (G (C_Carrot U Carrot_FarFrom_CT) & G F (Carrot_OnTopOf_CB & Carrot_CloseTo_CB) & X (G (C_Knife U Knife) & G (C_Knife U Knife_FarFrom_CT) & G (C_Knife U Knife_OnTopOf_Carrot) & G ((C_Carrot & Knife_CloseTo_Carrot) U Carrot_Pieces)))), where "X", "F", "U", and "G" are LTL operators that denote "neXt", "Finally", "Until" and "Globally (Always)" respectively; "C_Carrot", "Carrot_FarFrom_CT", "Carrot_CloseTo_CB", and "Carrot_Pieces" are grounded predicates that denote the relations "gripper is close to the carrot", "carrot is far from the counter top", "carrot is close to the cutting board", and "carrot state where it has been cut into pieces" respectively.

To the best of our knowledge, there is no work that learns a mapping from natural language sentences and gestures to an LTL formula. The closest work is (Wang et al. 2021), which addresses the problem of learning to translate a natural language sentence to an LTL formula. We propose to add a gesture encoder into this work together with the original language encoder. Here the predicted LTL formula from a pair of speech and gesture is more like a sub-task that a human is specifying to a robot. Based on a sequence of recognized sub-tasks, we could further do shallow domain based plan recognition (Zha et al. 2017; Zhuo et al. 2020), where a planning domain is learned from sequences of sub-tasks.

Conclusion

In this work, we introduce NatSGD, a multimodal HRI dataset with human commands in the form of unstructured speeches and natural gestures, and robot behavior in the form of synchronized demonstrated trajectories of robots. We detail our meticulous experiment designs for collecting natural human behavior while interacting with a robot to accomplish complex household tasks. We developed a photo-realistic simulator that enables researchers to conduct tele-operations. It also works as a learning platform for the robot. The NatSGD dataset consists of rich annotations that facilitate the learning of multiple downstream household tasks.

Acknowledgments

We would like to thank peers and faculties from the UMD CS department and the Perception and Robotics Group for their valuable feedback and discussions. Special thanks to Lindsay Little, Dr. Vibha Sazawal, Dr. Michelle Mazurek, Nirat Saini, Virginia Choi, Dr. Chethan Parameshwara, and Dr. Nitin Sanket. We want to thank and recognize the contributions of Aavash Thapa, Sushant Tamrakar, Jordan Woo, Noah Bathras, Zaryab Bhatti, Youming Zhang, Jiawei Shi, Zhuoni Jie, Tianpei Gu, Nathaneal Brain, Jiejun Zhang, Daniel Arthur, Shaurya Srivastava, and Steve Clausen. Without their contributions and support, this work would not have been possible. Finally, the support of NSF under grant OISE 2020624 is greatly acknowledged.

References

Abeyruwan, S.; Graesser, L.; D'Ambrosio, D. B.; Singh, A.; Shankar, A.; Bewley, A.; and Sanketi, P. R. 2022. i-sim2real: Reinforcement learning of robotic policies in tight human-robot interaction loops. *arXiv preprint arXiv:2207.06572*.

Abner, N.; Cooperrider, K.; and Goldin-Meadow, S. 2015. Gesture for linguists: A handy primer. *Lang. Linguist. Compass*, 9(11): 437–451.

Ahuja, C.; Lee, D. W.; Ishii, R.; and Morency, L.-P. 2020. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Findings of the association for computational linguistics: EMNLP 2020*, 1884–1895.

Azagra, P.; Golemo, F.; Mollard, Y.; Lopes, M.; Civera, J.; and Murillo, A. C. 2017. A multimodal dataset for object model learning from natural human-robot interaction. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6134–6141. IEEE.

Banaji, M. R.; and Greenwald, A. G. 2013. *Blindspot: Hidden biases of good people*. Bantam.

Baran, R.; Tan, X.; Varnai, P.; Yu, P.; Ahlberg, S.; Guo, M.; Cortez, W. S.; and Dimarogonas, D. V. 2021. A ROS Package for Human-In-the-Loop Planning and Control under Linear Temporal Logic Tasks. In *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, 2182–2187. IEEE.

Bischoff, M. 2021. ros-sharp: ROS# is a set of open source software libraries and tools in C# for communicating with ROS from .NET applications, in particular Unity3D.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1): 172–186.

Cauchard, J. R.; E, J. L.; Zhai, K. Y.; and Landay, J. A. 2015. Drone & me: an exploration into natural human-drone interaction. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, 361–365. New York, NY, USA: Association for Computing Machinery.

Chang, J.-Y.; Tejero-de Pablos, A.; and Harada, T. 2019. Improved optical flow for gesture-based human-robot interaction. In *2019 International Conference on Robotics and Automation (ICRA)*, 7983–7989. IEEE.

Charbonneau, E.; Miller, A.; and LaViola, J. J. 2011. Teach me to dance: exploring player experience and performance in full body dance games. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*, number Article 43 in ACE '11, 1–8. New York, NY, USA: Association for Computing Machinery.

Chen, F.; Lv, H.; Pang, Z.; Zhang, J.; Hou, Y.; Gu, Y.; Yang, H.; and Yang, G. 2018. WristCam: A wearable sensor for hand trajectory gesture recognition and intelligent human–robot interaction. *IEEE Sensors Journal*, 19(19): 8441–8451.

Chen, H.; Leu, M. C.; and Yin, Z. 2022. Real-Time Multi-modal Human-Robot Collaboration Using Gestures and Speech. *Journal of Manufacturing Science and Engineering*, 1–22.

Codevilla, F.; Müller, M.; López, A.; Koltun, V.; and Dosovitskiy, A. 2018. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, 4693–4700. IEEE.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Dahlbäck, N.; Jönsson, A.; and Ahrenberg, L. 1993. Wizard of Oz studies. In *Proceedings of the 1st international conference on Intelligent user interfaces - IUI '93*. New York, New York, USA: ACM Press.

de Wit, J.; Krahmer, E.; and Vogt, P. 2021. Introducing the NEMO-Lowlands iconic gesture dataset, collected through a gameful human–robot interaction. *Behavior Research Methods*, 53(3): 1353–1370.

FFmpeg.org. 2021. FFmpeg Hardware Acceleration Introduction. https://trac.ffmpeg.org/wiki/HWAccelIntro. Accessed: 2021-9-29.

Finucane, C.; Jing, G.; and Kress-Gazit, H. 2010. LTLMoP: Experimenting with language, temporal logic and robot control. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1988–1993. IEEE.

Fothergill, S.; Mentis, H.; Kohli, P.; and Nowozin, S. 2012. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1737–1746. New York, NY, USA: Association for Computing Machinery.

Furgale, P.; Rehder, J.; and Siegwart, R. 2013. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1280–1286. IEEE.

Gomez Chavez, A.; Ranieri, A.; Chiarella, D.; Zereik, E.; Babić, A.; and Birk, A. 2019. CADDY underwater stereo-vision dataset for human–robot interaction (HRI) in the context of diver activities. *Journal of Marine Science and Engineering*, 7(1): 16.

Guo, M.; Johansson, K. H.; and Dimarogonas, D. V. 2013. Revising motion planning under linear temporal logic specifications in partially known workspaces. In *2013 IEEE international conference on robotics and automation*, 5025–5032. IEEE.

Habibie, I.; Elgharib, M.; Sarkar, K.; Abdullah, A.; Nyatsanga, S.; Neff, M.; and Theobalt, C. 2022. A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech. In *Special Interest Group on Computer*

Graphics and Interactive Techniques Conference Proceedings, 1–9.

Habibie, I.; Xu, W.; Mehta, D.; Liu, L.; Seidel, H.-P.; Pons-Moll, G.; Elgharib, M.; and Theobalt, C. 2021. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 101–108.

Hart, S. G. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, 904–908. Sage publications Sage CA: Los Angeles, CA.

James, J.; Tian, L.; and Watson, C. I. 2018. An open source emotional speech corpus for human robot interaction applications. In *Interspeech*, 2768–2772.

Kalegina, A.; Schroeder, G.; Allchin, A.; Berlin, K.; and Cakmak, M. 2018. Characterizing the Design Space of Rendered Robot Faces. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, 96–104. New York, NY, USA: Association for Computing Machinery.

Kaspar, M.; Osorio, J. D. M.; and Bock, J. 2020. Sim2real transfer for reinforcement learning without dynamics randomization. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4383–4388. IEEE.

Kesten, Y.; Pnueli, A.; and Raviv, L.-o. 1998. Algorithmic verification of linear temporal logic specifications. In *International Colloquium on Automata, Languages, and Programming*, 1–16. Springer.

Konur, S. 2013. A survey on temporal logics for specifying and verifying real-time systems. *Frontiers of Computer Science*, 7(3): 370–403.

Krishnaswamy, N.; and Pustejovsky, J. 2020. A Formal Analysis of Multimodal Referring Strategies Under Common Ground. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 5919–5927.

Kucherenko, T.; Jonell, P.; Yoon, Y.; Wolfert, P.; and Henter, G. E. 2021a. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In *26th international conference on intelligent user interfaces*, 11–21.

Kucherenko, T.; Nagy, R.; Neff, M.; Kjellström, H.; and Henter, G. E. 2021b. Multimodal analysis of the predictability of hand-gesture properties. *arXiv preprint arXiv:2108.05762*.

Lee, G.; Deng, Z.; Ma, S.; Shiratori, T.; Srinivasa, S. S.; and Sheikh, Y. 2019. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 763–772.

Liu, X.; Shi, H.; Chen, H.; Yu, Z.; Li, X.; and Zhao, G. 2021. iMiGUE: An Identity-Free Video Dataset for Micro-Gesture Understanding and Emotion Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10631–10642. openaccess.thecvf.com.

Luan, W.; Yang, Y.; Fermüller, C.; and Baras, J. S. 2016. Reliable attribute-based object recognition using high predictive value classifiers. In *European Conference on Computer Vision*, 801–815. Springer.

Lyons, J. 1977. *Semantics: Volume 1*. ACLS Humanities E-Book. Cambridge University Press. ISBN 9780521291651.

Matuszek, C.; Bo, L.; Zettlemoyer, L.; and Fox, D. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Mazhar, O.; Ramdani, S.; Navarro, B.; Passama, R.; and Cherubini, A. 2018. Towards real-time physical human-robot interaction using skeleton information and hand gestures. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1–6. IEEE.

Narayan-Chen, A.; Jayannavar, P.; and Hockenmaier, J. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5405–5415.

Neto, P.; Simão, M.; Mendes, N.; and Safeea, M. 2019. Gesture-based human-robot interaction for human assistance in manufacturing. *The International Journal of Advanced Manufacturing Technology*, 101(1): 119–135.

Novoa, J.; Escudero, J. P.; Fredes, J.; Wuth, J.; Mahu, R.; and Yoma, N. B. 2017. Multichannel robot speech recognition database: MChRSR. *arXiv preprint arXiv:1801.00061*.

Nuzzi, C.; Pasinetti, S.; Pagani, R.; Coffetti, G.; and Sansoni, G. 2021. HANDS: an RGB-D dataset of static hand-gestures for human-robot interaction. *Data in Brief*, 35: 106791.

Otter.ai. 2021. Otter: Automatic Speech Recognition Tool. https://otter.ai/. Accessed: 2021-9-29.

Padmakumar, A.; Thomason, J.; Shrivastava, A.; Lange, P.; Narayan-Chen, A.; Gella, S.; Piramuthu, R.; Tur, G.; and Hakkani-Tur, D. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2017–2025.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pisharady, P. K.; and Saerbeck, M. 2015. Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 141: 152–165.

Pnueli, A. 1977. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, 46–57. ieee.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Rodomagoulakis, I.; Kardaris, N.; Pitsikalis, V.; Arvanitakis, A.; and Maragos, P. 2016. A multimedia gesture dataset for human robot communication: Acquisition, tools and recognition results. In *2016 IEEE International Conference on Image Processing (ICIP)*, 3066–3070. IEEE.

Shrestha, S.; Sentosatio, W.; Peng, H.; Fermuller, C.; and Aloimonos, Y. 2023. FEVA: Fast Event Video Annotation Tool. *arXiv preprint arXiv:2301.00482*.

Shukla, D.; Erkent, Ö.; and Piater, J. 2016. A multi-view hand gesture rgb-d dataset for human-robot interaction scenarios. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, 1084–1091. IEEE.

Starke, S.; Hendrich, N.; Krupke, D.; and others. 2017. Evolutionary multi-objective inverse kinematics on highly articulated and humanoid robots. *2017 IEEE/RSJ*.

Starke, S.; Hendrich, N.; and Zhang, J. 2018. Memetic evolution for generic full-body inverse kinematics in robotics and animation. *IEEE Transactions on*.

Stepputtis, S.; Campbell, J.; Phielipp, M.; Lee, S.; Baral, C.; and Amor, H. B. 2020. Language-Conditioned Imitation Learning for Robot Manipulation Tasks. arXiv:2010.12083.

Takashima, K.; Omori, Y.; Yoshimoto, Y.; Itoh, Y.; Kitamura, Y.; and Kishino, F. 2008. Effects of avatar's blinking animation on person impressions. In *Graphics Interface*, 169–176. researchgate.net.

Thomas, F.; Johnston, O.; and Thomas, F. 1995. *The illusion of life: Disney animation*. Hyperion New York.

Tomasello, M. 2010. *Origins of human communication*. MIT press.

Trutoiu, L. C.; Carter, E. J.; Matthews, I.; and Hodgins, J. K. 2011. Modeling and animating eye blinks. *ACM Trans. Appl. Percept.*, 8(3): 1–17.

Unity Technologies. 2021. Unity 3D Game Engine. https://unity.com/. Accessed: 2021-9-28.

Vasudevan, A. B.; Dai, D.; and Van Gool, L. 2018. Object referring in visual scene with spoken language. In *2018 IEEE winter conference on applications of computer vision (WACV)*, 1861–1870. IEEE.

Wang, C.; Ross, C.; Kuo, Y.-L.; Katz, B.; and Barbu, A. 2021. Learning a natural-language to LTL executable semantic parser for grounded robotics. In *Conference on Robot Learning*, 1706–1718. PMLR.

Wikipedia contributors. 2021. Baxter (name). https://en.wikipedia.org/wiki/Baxter_(name). Accessed: NA-NA-NA.

Yang, Y.; Li, Y.; Fermuller, C.; and Aloimonos, Y. 2015. Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Zha, Y.; Li, Y.; Gopalakrishnan, S.; Li, B.; and Kambhampati, S. 2017. Recognizing plans by learning embeddings from observed action distributions. *arXiv preprint arXiv:1712.01949*.

Zhuo, H. H.; Zha, Y.; Kambhampati, S.; and Tian, X. 2020. Discovering underlying plans based on shallow models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(2): 1–30.

Appendix

Dataset Collection Additional Details

Participants Eighteen volunteers participated in this experiment (9 male and 9 female) and received cash of $10 each. Participants were recruited through posting flyers around the university and using local mailing lists. Participants' ages ranged from 18 to 31 years (Mean 20.91±3.75). Their personalities were generally characterized as follows: Extroversion (5.56±2.09), Agreeableness (9.17±1.15), Conscientiousness (8.17±1.38), Emotional Stability (7.56±1.79), and Openness to Experience (7.89±1.47). Equal distribution of technical to non-technical background participants were chosen based on our questionnaire and post interview where we assessed their exposure to robots, remote controlled or gesture controlled games or toys, and their whether their education or profession was considered to have a technology focus such as computer science and engineering. None of the participants had ever interacted with any type of robots before they participated in the study.

Lab Setup Participants were invited to the lab where we showed the Baxter humanoid and video recording of the robot performing kitchen actions that demonstrated the robot's ability to fetch items from a refrigerator, microwave a dish, and prepare a salad (see Fig. 6). In addition, an we showed another video of an interaction between Baxter and a human where Baxter is following commands given by pointing or other demonstrations made by the body. We explained that since these videos, we have made a lot of improvements and wanted to understand how well the robot works with people and how well people can work with the robot. To be more believable, we explained that the robot motors were slow, so for the timing constraint, we created a virtual version of the robot that has the same AI engine. This allowed the participants to interact with virtual Baxter through a large monitor (55") at a 7-10' distance as shown in Fig. 7. Data was recorded from three Stereolabs ZED cameras from the front (camera 1), left and right at at approximately 30 degrees angle, at a resolution of 1280×720 px. (720p) at 30 frames per second (FPS). A single iPhone 6S also recorded the participants audio and video in portrait mode at at a resolution of 720p at 30 FPS from approximately 15 degrees angle.

Experiment Procedures *i) Signing Up* Participant fill out a screening questionnaire to sign up. We collect information about their demographics, contact info for scheduling, language proficiency, and dominant hand information.

ii) Briefing Participants are welcomed and given a short introduction of the lab and the work we are doing. They then sign consent forms if they wish to continue. We show recording of the past demonstrations of the humanoid performing various kitchen tasks. They are explained that due to the slow movement of the physical robot, they will be interacting with a simulated version with the same artificial intelligent brain.

iii) Simulator As shown in Fig. 2, the NatSGD Robot Simulator was built using Unity 3D (Unity Technologies 2021) with modified ROS plugin (Bischoff 2021). The system runs on a Intel i7 Gen 16GB RAM connected to the 55" TV. On the top right corner of the screen, a camera feed of the participant is overlayed as shown in Fig 1. This serves as a feedback mechanism to the participants to help them stay within the frame. The simulation runs based on ROS storyline module that we developed that helps us design the workflow of the recipe in the Unity environment. ROS messaging can (a) control camera angles for staging the desired point of view of the scene and the robot (b) the video instructions of the step of the recipe (c) robot responses such as head nodding, confusion face, and normal blinking face, and (d) robot actions such as navigation or performing cooking tasks. The robot navigation path was predetermined as there were limited number of target locations while the realtime inverse kinematics (IK) of the head nods and robot arms were implemented with BioIK (Starke et al. 2017; Starke, Hendrich, and Zhang 2018). To wait for participant command, the robot looks at the participant. Before performing each task, the robot looks at the target object(s) to demonstrate robot's attention.

iv) Practice Session During the practice session, participants interact with the robot to navigate the robot in discrete steps in the kitchen and perform steps to cut an apple. The practice session lasts roughly 5 minutes until the participant feels comfortable. During practice phase, the participants can interact with the research facilitator. After the practice phase, once the data collection begins, they are not allowed to ask any questions until the data collection session was over.

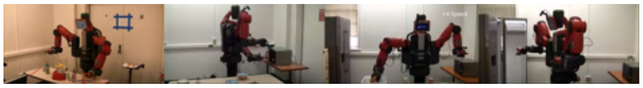


Figure 6: Video Demonstration of the robot performing kitchen tasks (a) mixing drinks (b) microwaving bowl (c) cleaning up (d) fetch milk from fridge.

v) WoZ Control Policy For both practice and the experiment sessions, the researcher facilitator (wizard) controls the workflow to move ahead as long as the participant's command is related to the task at hand and is discernible. Robot nods "yes" (head up and down motion) on commands that are understood. For unintelligible or unrelated commands, the robot displays the confusion face. For example, if a participant gestures to move right, or say move right, the wizard makes the robot move right. However, if the participant mumbles and the wizard cannot hear the participant, the wizard prompts with a confused robot face for clarification from the participant. The clarification questions are designed to be ambiguous so the participant attempts different strategies in providing instructions for the same task. The participant is allowed a maximum of five attempts per command. If unsuccessful after that, the task is skipped. In rare cases, the task can also be skipped, if the Unity game becomes frozen or Baxter is not able to solve the IK for grasping objects for more than 30 seconds.

vi) Post Survey and Interview After the data collection, participants are asked to fill out a survey providing their impression of the robot, their own personality, and the workload experienced. We finally wrap up the session with a semi-structured interview gathering additional details about their experience and debriefing them of any information we shared that was used to deceive them.

Background Noise One hypothesis was that background noise can cause people to use more gestures. We considered 3 types of noise recording playback (lawn mower, people talking, and music), but only tested with people talking as background noise as that was the only example people found to be believable and not simulated. We tested at 3 sets of loudness (M (dB) = 58, 63, 70, SD = 10, 13, 15). In our study (N=8), from people's use of speech and gesture and the post-interview, we found that (a) people tune out the background noise, instead of using more gestures or, (b) people wait for gaps of silence or lower level noise in cases of speech or periodic noise, and (c) the noise had to be so loud that none of the speech can be heard at all for them to use gestures instead of speech. For these reasons, we decided not to use background noise as an independent variable.

Robot Face and Name To reduce the affect of perceived gender, age, and personality by manipulating facial attributes, we considered the 17 face dimensions based on (Kalegina et al. 2018) study to design the face of the robot to be the most neutral face. The mouth of the robot was removed as not having a mouth did not have significant adverse effect on the neutral perception of the robot. Having a mouth seemed to give people the idea that the robot could speak, potentially causes the participant to prefer speech over gesture. For the robot to appear dynamic, friendly, and intelligent, we made the robot blink randomly between 12 and 18 blinks per minute (Takashima et al. 2008) with ease-in and ease-out motion profile (Trutoiu et al. 2011; Thomas, Johnston, and Thomas 1995). We further conducted pilot tests to analyze the head nod motions (velocity and number of nods) and facial expressions for confusion expression. Additionally, we avoided using gender specific pronouns "he/him" and "she/her" and referred to the robot as "the robot" or "Baxter" which is also the manufacturer given name printed on robot body that tends to be used both as a male and female name (Wikipedia contributors 2021).

Practice Session During practice, it is important to make sure that participants are not primed to use one modality versus the other. So steps were taken to design the session with a mixture of related and unrelated commands where both speech and gestures were used to command the robots. If participants used a single modality only, they were encouraged to test out using the other modality. Participants interacted with the robot and asked researchers questions during practice. Once the practice was completed, participants were not allowed to interact with anyone other than the robot even if they had questions or felt stuck as they were told that the experiment was designed for them to experience such scenarios and had to use creative methods make the robot understand what they wanted the robot to do.

WoZ Clues People can be quite intuitive in figuring out the patterns such as key press and mouse click sounds corresponding to robot actions. We experimented with masking the actual clicks and key presses with random ones. However, in the post interview the pilot test participants still seem to be able to figure out that researchers might be controlling the robot. So we created a soft rubber remote control keys that use IR receiver using Arduino micro-controller USB adapter to send keys to the WoZ UI with virtually no sound that the researcher keeps in their pocket. With this implementation, during the experiment, the researchers made sure when the experiment is being conducted, they do not sit at the control computer and appear to be moving around doing other things appearing busy, staring at their phone seemingly distracted, or looking at the participants showing attention in making sure the system was working without any technical issues. With this implementation, 100% of the participants believed that the robot was acting on its own and none of the participants suspected the WoZ setup to be a possibility.

Instructions Based on the recommendations (Fothergill et al. 2012), we tested various modalities for our applications. Our findings in our pilot studies were in-line with (Fothergill et al. 2012; Charbonneau, Miller, and LaViola 2011) where the instruction modality had a significant impact on the participants' behavior. For instances where text instructions were provided similar to (Cauchard et al. 2015), participants preferred speech and used the exact words for the action and the object with little or no gestures. With videos of people performing the task similar to (Charbonneau, Miller, and LaViola 2011), participants copied the exact style of the demonstration of the actor. The one with the most variance in speech vocabulary and styles of gestures were when we showed before-after video clips to show the pre-task and post-task states, for example, to turn on a stove, we showed a zoomed in video of a stove that was turned off, and faded out to a video of the stove with fire burning. For cutting apple, video of a whole apple on a cutting board being approached by a knife and faded into apple that was cut into pieces where the knife is leaves the screen. And these videos were repeated in a loop with a 1 second gap in between.

Story Line The experiment was organized for the participants to spontaneously communicate with a robot by commanding it to follow step-by-step start-to-finish instructions to prepare food in the kitchen. The food preparation steps were designed so that most of the tasks were repeated at least two times with different objects. The robot, in random order, alternated between performing the task immediately after the command was given and requiring the participants to provide additional instructions or disambiguate the original command. A vegetable soup recipe was designed for the experiment and included prepping, cooking, and serving steps. For prepping, the ingredients had to be fetched and cut. The cooking steps included turning a gas stove burner on and off, sauteing, stirring, seasoning, transferring, and covering/uncovering a pot. Serving the soup required fetching a bowl and pouring the soup into the bowl and placing it on the counter. The final steps included cleaning up by putting away the pots and pans into the sink.

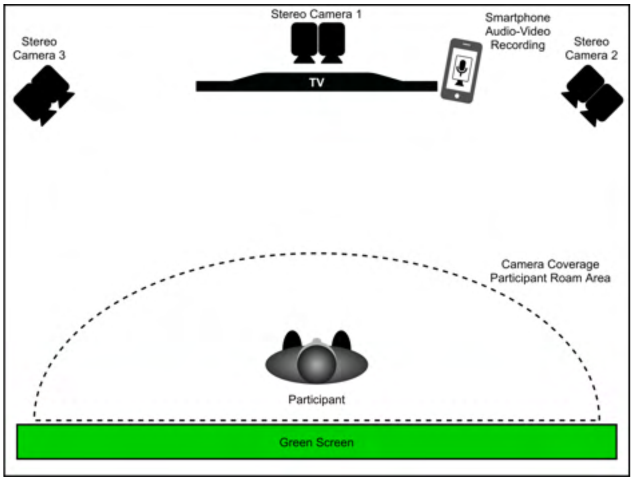


Figure 7: Experiment Layout top view and Panorama image of the lab setup

Dataset Structure and Labels The dataset structure is organized in the following way: For every instance of command there is an ID along with the following: corresponding IDs for *participant*, *video*, *speech*, *gesture*, *pose*, and *state*, time *onset* and *offset* of *speech* and *gesture*, text for *speech*, task group of *action object* tuple (more details in below), one-hot encoding of *speech* and reference of *speech* for *object* and *action*, one-hot encoding of *gesture*, reference of *gesture* for *object* and *action*, gesture containing *task* and *non-intent*, one-hot encoding for all 6 body parts for *task action gesture*, *task object gesture*, and *non-intent gesture*.

There are eleven (11) *Action* groups namely: *Add, Clean Up, Cut, Fetch, Put On, Serve, Stir, Take Off, Transfer, Turn Off,* and *Turn On* and twenty (20) interaction *Object* groups namely: *Carrot, Celery, Lid, Oil, Pan, Pepper, Pot, Potato, Salt, Soup, Spices, Tomato, Veges, Cutting Board, Knife, Bowl, Spatula, Ladle, Stove,* and *Sink*. As shown in Fig. 8, *Objects* can have various states for example, *Onion* can be whole or cut represented *Onion(Cut, Whole)* or *Onion(Cut, Pieces)*. Similarly, other attribute such as On, *In*, *Out*, *Covered*, *Uncovered*, *TurnedOn*, *TurnedOff*, *Contains*, *Visible*, NotVisible, *Grasped*, *Ungrasped*, and six degrees of freedom *Location* and *Rotation*. *Action*s or *Time* can causes the *Object* states to change. Since all the objects and actions exist in a single plane, participant *gaze* and *pointing* gesture can be in one of three directions *Left*, *Middle*, or *Right* shown my the black dotted lines in Fig. 8. *Video ID*, *pose ID*, and *state ID* all refer to external files in their corresponding folder. Data is available in two formats *CSV* and *Python dictionary pickle* file. Python scripts will be provided to load and parse each file.

Post Processing: Data Processing, Synchronization, and Camera Calibration It is important to clean up the data and make it easy for researchers to use the data. Careful iterative steps were taken to prepare the data to ensure integrity, quality, validity, and fairness. The raw data is processed, annotated, validated, visualized, and curated for downstream analysis and machine learning tasks in the following way.

i) Multi-camera Calibration: A standard 12×8 5” checker board was recorded using ROS, and Kalibr package (Furgale, Rehder, and Siegwart 2013) to compute the cameras intrinsic and extrinsic matrix. If the average re-projection error was greater than 1 px., the calibration was repeated.

ii) Multi-camera Audio-Video Synchronization and Data Compression: All the data was recorded using ROS bag. These recorded video frames from each cameras tend to have dynamically varying frames per second rates anywhere from 25 fps to 32 fps which makes it difficult to synchronize with sound. For this reason, the audio recording is extracted from a well established audio-video camera such as Apple iPhone camera. A flashing color screen from another computer is placed in the middle of the lab within all the cameras' field of view. A ROS start message is also published to store and identify the starting flag of the session for all other data. The changing color from red to blue is used to denote the mark of the starting frame and the ROS bag start message time is used for offsetting other messages. The frames are then streamed to a canvas that is 6× the size of 720p i.e., 2560×2560 where each row is a 720p stereo camera frame. At 33.33ms the latest state of the frame is recorded. The iPhone video is also clipped starting from the blue frames whose sound is then merged with the large canvas video to generate the data. This data is then re-encoded to be compressed using FFMPEG and NVIDIA TITAN X H.264 encoder (FFmpeg.org 2021).

Data Annotation Similar to (Liu et al. 2021), multiple people annotated the data for purposes of speed and reliability. Each annotations were annotated at least twice with up to three round of agreement checks. Data with difference in opinion that could not come to agreement were subject to voting by five annotators to require minimum 80% score, otherwise were left out from the final label list. The data was annotated using Fast Event Video Annotation Tool (FEVA (Shrestha et al. 2023)) (see Fig. 3). Speech and gesture onset and offset of each command is annotated by two researchers. The data is stored using a `json` format with FEVA dataset schema v2.1. Speech was generated using Otter (Otter.ai 2021) which was audited by two researchers. The human pose were extracted using OpenPose (Cao et al. 2019) which we filtered to discard frames with large errors or joints that were not detected. The full dataset schema is detailed in the appendix. Each event is then annotated by five independent annotators using FEVA crowd based on the ontology as shown in Fig. 5.

Inter-rater Reliability Inter-rate reliability (IRR) was computed using Cohen's Kappa (Cohen 1960) and annotation was updated in three rounds where 100% agreement was reached. The round 1 IIR for modality was 99.7% for speech, 94.4% for gesture. Similarly, for speech reference for object was 97.3%, for action was 99.3%. For gesture type, Task oriented was 85.2%, and non-intent was 72.0%. For gesture reference for object was 77.9%, and for action 80.3%. For task object reference body parts was 91.4%, task action body parts 88.8%, and non-intent body parts was 79.9%. All scoring well above recommended 70%. Modality and speech reference reached 100% agreement in round 2 correcting for mistakes and revised look. For gesture type, reference gesture, and gesture body parts, most reached approximately 90% in round 2 and 100% in round 3.

Dataset Statistical Insights

Insight 1 Overall 97.3% of the commands contained speech and 81.5% contained gesture, with 81.0% containing both at the same time. However, commands with speech but without gesture drops to 18.46% and gesture only to 2.7%. This implies much higher preference for speech over gesture, but also shows gesture being used with speech significantly.

Insight 2: While gestures are used 81.54% of the time, it is only used 2.7% of time implying people are less likely to use gesture independently in the context of kitchen tasks. When gestures are used, 74.68% of the time was conveying task oriented messages while 56.55% of the time also contained non-intent gestures. Independently, task oriented message that did not have any non-intent gestures were 42.70% of the time. This implies non-intent gestures can be confounding the overall gesture and wrongly interpreted by models that naively model the human motion. For instance, there are several instances where some participants use a rhythmic (beat) gestures to regulate and help them express a message they are having difficulty expressing. Those instances, naively the gesture look a cutting motion. Without speech these gestures are ambiguous and can be misleading.

As shown in Fig. 9 majority of the time, right and left hands are being for gestures which seems obvious. However, regardless of the handedness, in a two-tailed T-test across participants, people on average used left hand more to indicate action gestures (M=63.50, SD=18.87) compared to (M=25.44, SD=13.78) with the right hand t=6.90933, $p<0.00001$. This can be explained by the fact that, most of the actions with the left hands were for activities that was happening to the left side of the user implying spatial priority people give to express with gesture.

Insight 3: The data format that we collected can be used to train machine learning models to use multimodal data for many tasks such as command recognition, visual question and answering, temporal segmentation of people behavior, learning cascaded command intents, and so on. These can be useful for teaching robots to understand our intents of our commands better.

Insight 4: As stated in *Insight 2* and in the speech data, people often refer to spatial objects and locations, items referred to from previous commands, using the same words for two items such as pot for both the pan and the pot, using words like fire to mean the stove and so on. These kinds of information is required by the robot to make observation outside of the human command to understand the bigger context of the commands. This dataset take a step in that direction. But more work is needed.

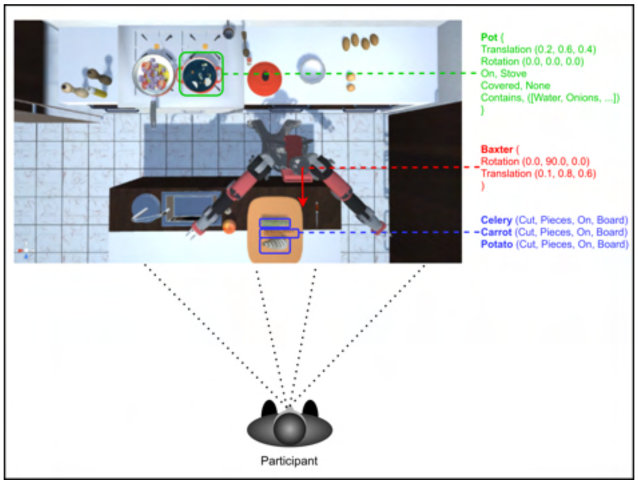

Pot {
Translation (0.2, 0.6, 0.4)
Rotation (0.0, 0.0, 0.0)
On, Stove
Covered, None
Contains, ([Water, Onions, ...])
}
Baxter {
Rotation (0.0, 90.0, 0.0)
Translation (0.1, 0.8, 0.6)
}
Celery (Cut, Pieces, On, Board)
Carrot (Cut, Pieces, On, Board)
Potato (Cut, Pieces, On, Board)
Participant

Figure 8: NatSGD Dataset (a) Participant Point of View (POV) first person view of the kitchen scene (b) Top view of the states of the kitchen, the robot, and the participant.

- *Modality*: The first distinction is the modality of the communications with the presence or absence of speech and gesture. Speech includes primarily the language, while para-linguistic features are reserved for future work. Gestures include full body static postures or dynamic movements, however, facial expressions and eye gaze are planned for future work.
- *Reference*: For robots to perform tasks, information of the *action* and the target *object* is key for the competition of the said task (Yang et al. 2015). Therefore, each *speech* and *gesture* clips are independently annotated for their references to the *objects* that the task requires the robot to interact with and the *actions* the robot needs to perform with or to the *objects*. These provide rich information on the "what", "where", and the "how" of the task.
- *Gesture Type*: Building on the Communicative-Informative dichotomy (Abner, Cooperrider, and Goldin-Meadow 2015) introduced by Lyons in Semantics (Lyons 1977), we segment the gestures based on the *intent* of the speaker as perceived by the receiver. We segment gestures into *Task* oriented and *Non-intent*. *Task* gestures serves to clearly and intentionally communicate the desired task to be performed. *Non-intent* gestures include static postures or dynamic movements that are not intended to communicate the specifics of the desired task. These gestures may provide additional contextual information, however, they have to be implied by the receiver and often tend to have a large degree of disagreement between annotators on what they mean or imply. Isolating these motions can help machine learning systems to classify gesture useful for interpreting the desired task information the speaker is intending to communicate from the gesture along with the speech. The *non-intent gestures* can also be used to gather the state of the speaker.
- *Body Parts*: For all task action or object or non-intent gestures, the corresponding body parts are clearly annotated and are divided into 6 body part groups: (1) *head* (2) *torso* (3) *left hand* (4) *right hand* (5) *left leg* and (6) *right leg*.

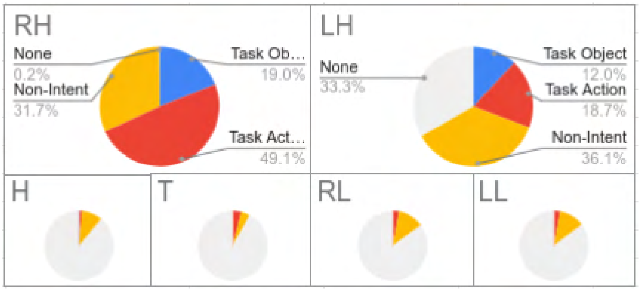


Figure 9: The distribution of gesture by body part usage for intent and reference dimensions for the gesture command. Right hand (RH), Left hand (LH), Head (H), Torso (T), Right leg (RL), and Left leg (LL).

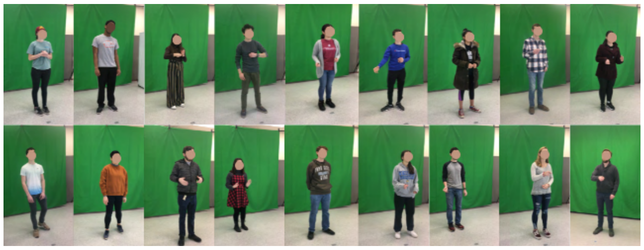


Figure 10: In this example, the participants are commanding the robot to cut onions where the participant's natural choice of communication is diverse.

References

Abeyruwan, S.; Graesser, L.; D'Ambrosio, D. B.; Singh, A.; Shankar, A.; Bewley, A.; and Sanketi, P. R. 2022. i-sim2real: Reinforcement learning of robotic policies in tight human-robot interaction loops. *arXiv preprint arXiv:2207.06572*.

Abner, N.; Cooperrider, K.; and Goldin-Meadow, S. 2015. Gesture for linguists: A handy primer. *Lang. Linguist. Compass*, 9(11): 437–451.

Ahuja, C.; Lee, D. W.; Ishii, R.; and Morency, L.-P. 2020. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Findings of the association for computational linguistics: EMNLP 2020*, 1884–1895.

Azagra, P.; Golemo, F.; Mollard, Y.; Lopes, M.; Civera, J.; and Murillo, A. C. 2017. A multimodal dataset for object model learning from natural human-robot interaction. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6134–6141. IEEE.

Banaji, M. R.; and Greenwald, A. G. 2013. *Blindspot: Hidden biases of good people*. Bantam.

Baran, R.; Tan, X.; Varnai, P.; Yu, P.; Ahlberg, S.; Guo, M.; Cortez, W. S.; and Dimarogonas, D. V. 2021. A ROS Package for Human-In-the-Loop Planning and Control under Linear Temporal Logic Tasks. In *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, 2182–2187. IEEE.

Bischoff, M. 2021. ros-sharp: ROS# is a set of open source software libraries and tools in C# for communicating with ROS from .NET applications, in particular Unity3D.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1): 172–186.

Cauchard, J. R.; E, J. L.; Zhai, K. Y.; and Landay, J. A. 2015. Drone & me: an exploration into natural human-drone interaction. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, 361–365. New York, NY, USA: Association for Computing Machinery.

Chang, J.-Y.; Tejero-de Pablos, A.; and Harada, T. 2019. Improved optical flow for gesture-based human-robot interaction. In *2019 International Conference on Robotics and Automation (ICRA)*, 7983–7989. IEEE.

Charbonneau, E.; Miller, A.; and LaViola, J. J. 2011. Teach me to dance: exploring player experience and performance in full body dance games. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*, number Article 43 in ACE '11, 1–8. New York, NY, USA: Association for Computing Machinery.

Chen, F.; Lv, H.; Pang, Z.; Zhang, J.; Hou, Y.; Gu, Y.; Yang, H.; and Yang, G. 2018. WristCam: A wearable sensor for hand trajectory gesture recognition and intelligent human–robot interaction. *IEEE Sensors Journal*, 19(19): 8441–8451.

Chen, H.; Leu, M. C.; and Yin, Z. 2022. Real-Time Multi-modal Human-Robot Collaboration Using Gestures and Speech. *Journal of Manufacturing Science and Engineering*, 1–22.

Codevilla, F.; Müller, M.; López, A.; Koltun, V.; and Dosovitskiy, A. 2018. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, 4693–4700. IEEE.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Dahlbäck, N.; Jönsson, A.; and Ahrenberg, L. 1993. Wizard of Oz studies. In *Proceedings of the 1st international conference on Intelligent user interfaces - IUI '93*. New York, New York, USA: ACM Press.

de Wit, J.; Krahmer, E.; and Vogt, P. 2021. Introducing the NEMO-Lowlands iconic gesture dataset, collected through a gameful human–robot interaction. *Behavior Research Methods*, 53(3): 1353–1370.

FFmpeg.org. 2021. FFmpeg Hardware Acceleration Introduction. https://trac.ffmpeg.org/wiki/HWAccelIntro. Accessed: 2021-9-29.

Finucane, C.; Jing, G.; and Kress-Gazit, H. 2010. LTLMoP: Experimenting with language, temporal logic and robot control. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1988–1993. IEEE.

Fothergill, S.; Mentis, H.; Kohli, P.; and Nowozin, S. 2012. Instructing people for training gestural interactive systems.

In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1737–1746. New York, NY, USA: Association for Computing Machinery.

Furgale, P.; Rehder, J.; and Siegwart, R. 2013. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1280–1286. IEEE.

Gomez Chavez, A.; Ranieri, A.; Chiarella, D.; Zereik, E.; Babić, A.; and Birk, A. 2019. CADDY underwater stereo-vision dataset for human–robot interaction (HRI) in the context of diver activities. *Journal of Marine Science and Engineering*, 7(1): 16.

Guo, M.; Johansson, K. H.; and Dimarogonas, D. V. 2013. Revising motion planning under linear temporal logic specifications in partially known workspaces. In *2013 IEEE international conference on robotics and automation*, 5025–5032. IEEE.

Habibie, I.; Elgharib, M.; Sarkar, K.; Abdullah, A.; Nyatsanga, S.; Neff, M.; and Theobalt, C. 2022. A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, 1–9.

Habibie, I.; Xu, W.; Mehta, D.; Liu, L.; Seidel, H.-P.; Pons-Moll, G.; Elgharib, M.; and Theobalt, C. 2021. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 101–108.

Hart, S. G. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, 904–908. Sage publications Sage CA: Los Angeles, CA.

James, J.; Tian, L.; and Watson, C. I. 2018. An open source emotional speech corpus for human robot interaction applications. In *Interspeech*, 2768–2772.

Kalegina, A.; Schroeder, G.; Allchin, A.; Berlin, K.; and Cakmak, M. 2018. Characterizing the Design Space of Rendered Robot Faces. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, 96–104. New York, NY, USA: Association for Computing Machinery.

Kaspar, M.; Osorio, J. D. M.; and Bock, J. 2020. Sim2real transfer for reinforcement learning without dynamics randomization. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4383–4388. IEEE.

Kesten, Y.; Pnueli, A.; and Raviv, L.-o. 1998. Algorithmic verification of linear temporal logic specifications. In *International Colloquium on Automata, Languages, and Programming*, 1–16. Springer.

Konur, S. 2013. A survey on temporal logics for specifying and verifying real-time systems. *Frontiers of Computer Science*, 7(3): 370–403.

Krishnaswamy, N.; and Pustejovsky, J. 2020. A Formal Analysis of Multimodal Referring Strategies Under Common Ground. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 5919–5927.

Kucherenko, T.; Jonell, P.; Yoon, Y.; Wolfert, P.; and Henter, G. E. 2021a. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In *26th international conference on intelligent user interfaces*, 11–21.

Kucherenko, T.; Nagy, R.; Neff, M.; Kjellström, H.; and Henter, G. E. 2021b. Multimodal analysis of the predictability of hand-gesture properties. *arXiv preprint arXiv:2108.05762*.

Lee, G.; Deng, Z.; Ma, S.; Shiratori, T.; Srinivasa, S. S.; and Sheikh, Y. 2019. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 763–772.

Liu, X.; Shi, H.; Chen, H.; Yu, Z.; Li, X.; and Zhao, G. 2021. iMiGUE: An Identity-Free Video Dataset for Micro-Gesture Understanding and Emotion Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10631–10642. openaccess.thecvf.com.

Luan, W.; Yang, Y.; Fermüller, C.; and Baras, J. S. 2016. Reliable attribute-based object recognition using high predictive value classifiers. In *European Conference on Computer Vision*, 801–815. Springer.

Lyons, J. 1977. *Semantics: Volume 1*. ACLS Humanities E-Book. Cambridge University Press. ISBN 9780521291651.

Matuszek, C.; Bo, L.; Zettlemoyer, L.; and Fox, D. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Mazhar, O.; Ramdani, S.; Navarro, B.; Passama, R.; and Cherubini, A. 2018. Towards real-time physical human-robot interaction using skeleton information and hand gestures. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1–6. IEEE.

Narayan-Chen, A.; Jayannavar, P.; and Hockenmaier, J. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5405–5415.

Neto, P.; Simão, M.; Mendes, N.; and Safeea, M. 2019. Gesture-based human-robot interaction for human assistance in manufacturing. *The International Journal of Advanced Manufacturing Technology*, 101(1): 119–135.

Novoa, J.; Escudero, J. P.; Fredes, J.; Wuth, J.; Mahu, R.; and Yoma, N. B. 2017. Multichannel robot speech recognition database: MChRSR. *arXiv preprint arXiv:1801.00061*.

Nuzzi, C.; Pasinetti, S.; Pagani, R.; Coffetti, G.; and Sansoni, G. 2021. HANDS: an RGB-D dataset of static hand-gestures for human-robot interaction. *Data in Brief*, 35: 106791.

Otter.ai. 2021. Otter: Automatic Speech Recognition Tool. https://otter.ai/. Accessed: 2021-9-29.

Padmakumar, A.; Thomason, J.; Shrivastava, A.; Lange, P.; Narayan-Chen, A.; Gella, S.; Piramuthu, R.; Tur, G.; and Hakkani-Tur, D. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2017–2025.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pisharady, P. K.; and Saerbeck, M. 2015. Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 141: 152–165.

Pnueli, A. 1977. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, 46–57. ieee.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Rodomagoulakis, I.; Kardaris, N.; Pitsikalis, V.; Arvanitakis, A.; and Maragos, P. 2016. A multimedia gesture dataset for human robot communication: Acquisition, tools and recognition results. In *2016 IEEE International Conference on Image Processing (ICIP)*, 3066–3070. IEEE.

Shrestha, S.; Sentosatio, W.; Peng, H.; Fermuller, C.; and Aloimonos, Y. 2023. FEVA: Fast Event Video Annotation Tool. *arXiv preprint arXiv:2301.00482*.

Shukla, D.; Erkent, Ö.; and Piater, J. 2016. A multi-view hand gesture rgb-d dataset for human-robot interaction scenarios. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, 1084–1091. IEEE.

Starke, S.; Hendrich, N.; Krupke, D.; and others. 2017. Evolutionary multi-objective inverse kinematics on highly articulated and humanoid robots. *2017 IEEE/RSJ*.

Starke, S.; Hendrich, N.; and Zhang, J. 2018. Memetic evolution for generic full-body inverse kinematics in robotics and animation. *IEEE Transactions on*.

Stepputtis, S.; Campbell, J.; Phielipp, M.; Lee, S.; Baral, C.; and Amor, H. B. 2020. Language-Conditioned Imitation Learning for Robot Manipulation Tasks. arXiv:2010.12083.

Takashima, K.; Omori, Y.; Yoshimoto, Y.; Itoh, Y.; Kitamura, Y.; and Kishino, F. 2008. Effects of avatar's blinking animation on person impressions. In *Graphics Interface*, 169–176. researchgate.net.

Thomas, F.; Johnston, O.; and Thomas, F. 1995. *The illusion of life: Disney animation*. Hyperion New York.

Tomasello, M. 2010. *Origins of human communication*. MIT press.

Trutoiu, L. C.; Carter, E. J.; Matthews, I.; and Hodgins, J. K. 2011. Modeling and animating eye blinks. *ACM Trans. Appl. Percept.*, 8(3): 1–17.

Unity Technologies. 2021. Unity 3D Game Engine. https://unity.com/. Accessed: 2021-9-28.

Vasudevan, A. B.; Dai, D.; and Van Gool, L. 2018. Object referring in visual scene with spoken language. In *2018 IEEE winter conference on applications of computer vision (WACV)*, 1861–1870. IEEE.

Wang, C.; Ross, C.; Kuo, Y.-L.; Katz, B.; and Barbu, A. 2021. Learning a natural-language to LTL executable semantic parser for grounded robotics. In *Conference on Robot Learning*, 1706–1718. PMLR.

Wikipedia contributors. 2021. Baxter (name). https://en.wikipedia.org/wiki/Baxter_(name). Accessed: NA-NA-NA.

Yang, Y.; Li, Y.; Fermuller, C.; and Aloimonos, Y. 2015. Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Zha, Y.; Li, Y.; Gopalakrishnan, S.; Li, B.; and Kambhampati, S. 2017. Recognizing plans by learning embeddings from observed action distributions. *arXiv preprint arXiv:1712.01949*.

Zhuo, H. H.; Zha, Y.; Kambhampati, S.; and Tian, X. 2020. Discovering underlying plans based on shallow models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(2): 1–30.