Automatic Generation of Robot Actions for Collaborative Tasks from Speech

Manizheh Zand*§, Krishna Kodur†§, Maria Kyrarini‡
Department of Electrical & Computer Engineering,
Santa Clara University,
Santa Clara, CA, USA
Email: *mzand@scu.edu, †kkodur@scu.edu, ‡mkyrarini@scu.edu

***Abstract*—Robots have the potential to assist people in daily tasks, such as cooking a meal. Communicating with the robots verbally and in an unstructured way is important, as spoken language is the main form of communication for humans. This paper proposes a novel framework that automatically generates robot actions from unstructured speech. The proposed framework was evaluated by collecting data from 15 participants preparing their meals while seating on a chair in a randomly disrupted environment. The system can identify and respond to a task sequence while the user may be engaged in unrelated conversations, even if the user's speech might be unstructured and grammatically incorrect. The accuracy of the proposed system is 98.6%, which is a very promising finding.**

***Index Terms*—Human-robot collaboration, Robot Action Generation, Natural Language Processing, Assistive Cooking, Speech**

I. Introduction

According to the Centers for Disease Control and Prevention [1], 61 million adults in the United States live with a disability, which may include mobility, cognition, hearing, vision, and self-care challenges. People with disabilities often need support to perform Activities of Daily Living (ADLs). Independent living gives purpose and meaning to a person's life which improves the person's confidence, self-esteem, and quality of life.

Robots have the potential to assist people with disabilities to get some degree of independence. For example, robots have been used to provide drinks to people with severe motor impairments [2], [3] or help them eat [4], [5]. However, meal preparation may be a challenging task for people with disabilities. A study [6] that interviewed 30 people with disabilities in New Jersey (USA) found that cooking is a very complex task as most kitchens are not accessible to people in wheelchairs or people who are blind due to structural barriers (e.g., the counter is too high, impossible-to-reach storage cupboards, narrow spaces, flat-screen interface for microwaves and ovens, etc.). Robotic systems could assist people who would like to cook by opening cupboards, picking and bringing the ingredients and the kitchen utensils, operating the appliance, etc., while individuals can focus on actual cooking. Therefore, considering cooking as a human-robot collaborative scenario [7] is beneficial for people with disabilities.

Another important aspect of human-robot collaboration is the method of interaction. Spoken language interaction is a natural way for humans to communicate with their companions. However, our ability to communicate with robots via speech is very limited and restrictive [8]. For example, it is expected that the person interacting with the robot will provide specific instructive words for robotic actions. A recent study [9] developed a Wizard of Oz prototyping method to investigate how humans would interact with an industrial robot via speech. The participants of the study were able to instruct the virtual robot to move cubes and make a pyramid with them. The findings of the study suggest a high preference for speech input and the automatic generation of robotic actions.

Fig. 1. Experimental Setup for Human-Robot Collaborative Cooking.

This paper presents a robotic framework that enables a robotic system to generate automatically a sequence of actions for a collaborative cooking scenario from spoken language, as shown in Fig. 1. The proposed framework was evaluated by collecting data from 15 individuals and, during the study, random distractions and interruptions were introduced while they were interacting with the robot. The presented work contributes to the research and development of assistive robots by proposing the following: (I) a language model that can identify

This material is based upon work supported partially supported by the National Science Foundation under Grant 2226165 and by Santa Clara University under University Research Grant.

§Equal contribution

the intended task-related instructions to the robot from an unstructured speech (including distractions and interruptions), and (II) automatic generation of a graph-based sequence of robotic actions from spoken language.

The rest of the paper is organized into the following sections: Section 2 discusses the related work, Section 3 the Experimental setup and data collection, Section 4 the proposed system, and Section 5 the experimental results. Finally, Section 7 concludes the work and provides future directions of research.

II. Related Work

Interpreting the task that is spoken to the robot is one of the critical aspects and yet challenging in a robotic assistive system, especially for assisting the user in managing their daily lives, such as cooking their meals. Unhelkar et al. [10] proposed "CommPlan", a computational framework that decides if, what, and when to communicate with the human during human-robot collaboration. The CommPlan was used for a meal preparation task where humans and robots communicate in making a sandwich. The CommPlan predefines a set of spoken words that humans and robots can use to communicate. However, predefining a set of words does not provide a natural way of communication for humans, as humans have the ability to express the same exact action with different words and sentences.

González et al. [11] propose using syntactic rule-based parsers for extracting key action words (e.g. "start", "stop", etc.) for natural human-robot interaction in an industrial setting. However, the syntactic rule-based parsers might not be efficient in grasping a complex sentence [12]. To alleviate this issue Choi et al. [12] used Generative Pre-trained Transformer 2 (GPT-2) [13] to understand the user intent in a more natural way. Choi et al. [12] in their paper used GPT-2 to understand and execute the high-level verbal commands to perform motion-planning tasks, such as pick & place or assembly for industrial robots. In their architecture, the GPT-2 receives an input string such as "Please begin assembly of the casing base", which is then converted into an instruction task string, and then the framework instructs the robot to begin assembling the casing base. Similarly, Li et al. [14] propose a system called "ToD4IR" that uses the dialogues between the human and the robot where the human instructs the robot to perform either of these four tasks: Go to a location, deliver an object from point A to Point B, work in assembly and relocate the objects in a scene. The authors benchmark the popular language models, such as GPT-2 and GPT-Neo to find out the task recognition accuracy is 86.6%. Li et al. [14] and Choi et al. [12] focus on manufacturing related tasks that are structured scenarios.

In contrast, the presented work focuses on a meal preparation scenario that is unstructured. Additionally, to make the scenario more realistic, we introduced distractions and interruptions, such as a person trying to chat with the robot's user, a dog barking, the robot's user talking on the phone, etc. To the best of our knowledge, this is the first framework that focuses on generating automatic robotic actions from unstructured spoken language and at the same time identifying the relevant task instructions from unrelated ones.

III. Experimental Setup and Data Collection

The focus of our experimental setup is to understand how people would talk to a robot in order for it to pick and bring ingredients and kitchen utensils for cooking a meal of their interest. The setup includes two tables; one table has all the ingredients and kitchen utensils while the other table has a chair where the study's participant sits. The setup includes a microphone that records the interactions at the two tables. Fig 1 shows the experimental setup for this study. One of the study personnel pretends that s/he is the robotic assistant and follows the participant's instructions, while another study personnel interrupts and talks with the participant during an instruction or playing sounds (e.g. dog barking). Participants were allowed to answer their phone calls and make phone calls, as we were interested to collect data that are realistic. A mobile manipulator was also present in the room so the participants had a better understanding of what a robotic assistant looks like and its capabilities. Additionally, the participants were instructed to only use speech as a communication method and to imagine they were talking to a robot. Therefore, the participants were instructed to say "Hey robot" and then the action they would expect the robot to take. The participants were not instructed on how to provide the desired robot action. Each participant had the option to provide instructions for up to three meals of their choice, based on the available ingredients.

To collect data, 15 participants were selected, and Table I shows the participant's age, gender, and if they have familiarity with robots. The audio was recorded during the studies in accordance with Institutional Review Board (IRB) [15] Protocol ID 22-08-1827 and the described experimental setup was followed. The audio was recorded in ROSBags [16], a file format in Robot Operating System (ROS) for storing data. During the data collection, all the participants provided instructions to the robot for assisting them to prepare a meal. By not restricting the way the participants would talk to a robot, we observed the following behaviors. Fourteen participants used filler words (e.g. actually, literally, like I said, you know, what I am trying to say) or sounds during the study. Twelve participants mentioned the name of the recipe (e.g. pasta with tomato sauce, carbonara, etc.) at the start of each meal preparation. Moreover, six participants thought that the meal preparation was completed but then realized that they would like to add additional steps. Four participants requested the robot to track the time and three participants forgot the name of the ingredients or the utensils they needed. Only one participant thanked the robot when giving commands. The collected dataset from the participants has been named the *"cooking assistance"* dataset for further reference throughout the paper.

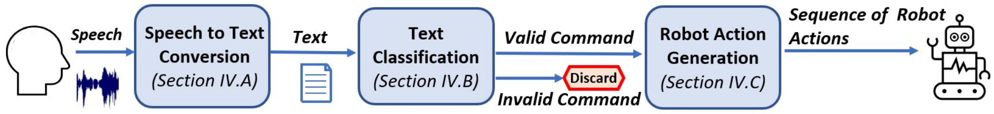

Fig. 2. Overview of the automatic generation of robot actions framework for collaborative tasks from speech

TABLE I
PARTICIPANTS' DEMOGRAPHICS: AGE GROUP, GENDER OF THE PARTICIPANTS, AND IF THEY ARE FAMILIAR WITH ROBOTS

Age Range	Gender (M: Male, F: Female)	Number of participants familiar with Robots
18-30	9M 1F	6
31-40	1M 1F	2
51-60	1M 1F	2
$\geq$60	1M 0F	0

IV. PROPOSED SYSTEM

The proposed system is shown in Figure 2, which consists of four modules described in detail in the following subsections.

A. Speech to Text Conversion

The first module of our system converts the speech (audio) data into text. There are two options for transcribing the audio to text; (1) online services from Google or Amazon [17] and (2) offline models, such as Vosk [18]. As Vosk gives the capability for offline transcription and runs locally on our server ensuring data privacy according to the IRB guidelines, it was preferred in comparison with online speech-to-text services. Vosk provides multiple trained models which can convert audio from 20+ languages to text. Out of all the models available, the Vosk English language model called the "Vosk-model-en-us-daanzu-20200905" [18] was chosen since it was trained to transcript speech from both native and non-native English speakers. The output of this module is the text transcript of the spoken language, which is further processed by the text classification module.

B. Text Classification

As we set up our study to be realistic, there is a significant chance that numerous conversations may not be relevant to meal preparation. For example, the user may get interrupted by other people and start a conversation with them or answer their phone. Therefore, it is vital that the system can distinguish between robot commands and unrelated conversations. One method that can classify if the text is a valid command or not is the open-source Bidirectional Encoder Representation from Transformers (BERT) [19], which is created by google in 2018. The data collected from the 15 participants is not enough to train a BERT model as it requires a large amount of data, such as the BookCorpus dataset [20], which contains text data from 11,038 unpublished books. To overcome this issue, the relevant data which is the valid commands given to the robot saying 'Hey robot', are augmented to increase its size. The data are augmented with random food names (e.g. Pizza, Steak, etc.) from [21], random food adjectives such as spicy and savory, etc. and also with different ingredients such as tomato and butter, etc. and different utensil names such as pan, pot, and different appliance names such as oven, toaster, etc. For example, if the participant says "I want to make pasta" during the data collection, this sentence is modified to "I want to make spicy and savory Pasta" and added to the *cooking assistance dataset*. From the data collected it can be inferred that there were two main types of commands that the participants asked the robot: (I) To fetch objects and (II) To set a timer.

To train the model to classify the commands intended for the robot as 'valid command', the collected valid commands from randomly chosen ten participants are augmented. Therefore, the *cooking assistance dataset* contains 10K examples of valid commands to the robot. Moreover, during the collection of the *cooking assistance dataset*, the participants were distracted while giving commands to the robot. There were cases where the participant said 'Hey robot' and went on to have a phone call rather than giving the command to the robot. The data collected during interruptions are a small sample and are not enough to train the model as examples of 'invalid command' for the robot. However, these examples of casual/distracted conversations can be easily obtained from other datasets, such as TweetQA [22]. The TweetQA is a dataset of informal human conversations and contains 13,757 question-and-answer pairs sampled from 17,794 tweets. Out of which 10,692 are used for training and 3065 are for testing. Both the questions and tweets from the TweetQA dataset are used to train our BERT model as examples of 'invalid command'.

The Binary Cross Entropy loss function is used to train the BERT model with our dataset. Let p_k be the output class of the model and y_k be the target class, then the loss function L can be defined as follows:

$$L = -\sum_{k=0}^{k=n} y_k log(p_k) \tag{1}$$

where n is the total number of examples in the dataset, which is around 20,692 samples (10,000 from the *cooking assistance dataset* and 10,692 from TweetQA). The model is trained for 20 epochs with a learning rate of 10^{-5} and weight decay 0.01 using Adam with weight decay optimizer [23]. The classified invalid commands are discarded, while the valid commands are processed further by the next module.

C. Robot Action Generation

The next step in our system is to generate a sequence of robot actions for the task based on the commands that have been classified as valid by the text classification module. The sequence of robot actions is organized as a graph and each robot action is a graph node. The edge of the graph describes which step it is (e.g. step 1, step 2, etc.).

The robot action generation module utilizes the causal language model, called Distilled Generative Pre-trained Transformer-2 (DistilGPT-2) [24], which is used to extract the important words from the valid command (e.g. if a user says "Hey robot, can you bring me the tomato", DistilGPT-2 will extract the words "bring" and "tomato"). As mentioned in the earlier subsection IV-B, a large amount of data is required to train language models. Hence, the same augmented dataset with just the valid commands that trained the BERT model is used to finetune the DistilGPT-2 model, which is pretrained with the WebText Dataset [13] from Huggingface [24].

Consider the input sentence S, which represents a valid command. S needs to be separated (tokenized) into individual words. Considering the tokenization yielded i words (tokens) $S = S_0,\ S_1,\ S_2,...,S_i$. Let the maximum number of tokens possible be m. The tokens are passed into the DistilGPT-2 model θ to generate the robot action R which contains m tokens, as follows:

$$R = \theta(S_0, S_1, S_2, \cdots, S_i) \tag{2}$$

Let the ground truth of the robot action be T which contains m tokens, then the total loss L used to train the DistilGPT-2 model is calculated using the following cross-entropy loss:

$$L = -\sum_{j=0}^{j=m} T_j log(R_j) \tag{3}$$

The generated robot actions from the DistilGPT-2 model are then converted into graph nodes and added to the graph of the task. Figure 3 shows an example of the participant's speech and the graph generated from our system to describe the sequence of robot actions for the desired task. Most participants provided the task's name (e.g. Let's make pasta with tomato sauce) when they start instructing the robot. However, three participants forgot to give the name of a task. In this case, we created an "unknown task" graph. After the participant provided all the instructions, then the graph includes the sequence of all robot actions required for the specific task. In the future, this graph can be used to directly control a mobile manipulator that is capable to recognize objects in the scene and fetch them for the user.

Task: Pasta Carbonara

Participant: Hey robot today I will be making pasta carbonara and I need a saucepan, olive oil and pasta
DistilGPT-2 Output: Add to Graph: pasta carbonara, Step-1 Fetch Saucepan, Step-2 Fetch olive oil, Step-3 Fetch pasta

Participant: Hey robot I need garlic bacon egg and cream
DistilGPT-2 Output: Add to current graph: bacon, Add to current graph: Fetch egg, Add to current graph: Fetch cream

Participant: Hey robot And then I need the chopping board and knife
DistilGPT-2 Output: Add to current graph: Fetch chopping board, Add to current graph: Fetch knife

Participant: Hey robot I need also the spatula
DistilGPT-2 Output: Add to current graph: Fetch spatula

Participant's Friend: Hey maria what are you making today, smells delicious
System: Invalid command. Text not sent to DistilGPT-2 for processing

Participant: Well I am trying to make pasta, you are more than welcome to join me
System: Invalid command. Text not sent to DistilGPT-2 for processing

Participant: Hey robot I just realized that I forgot the black pepper and oregano can you get them for me
DistilGPT-2 Output: Add to current graph: Fetch black pepper, Add to current graph: Fetch oregano

Participant: Hey robot I think you should take a break for until like 10 minutes and then we can continue the recipe
DistilGPT-2 Output: Add to current graph: Set Timer 10 minutes

Participant: Hey robot I need the strainer so that I can takeout the pasta from the pot
DistilGPT-2 Output: Add to current graph: Fetch strainer

Participant: Hey robot Thank you robot for the help we are finished
DistilGPT-2 Output: Done

Corresponding Graph:

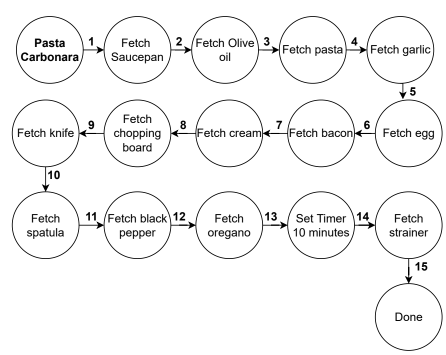

Fig. 3. Transcript of the speech given by a participant during the study and the corresponding generated graph for the sequence of robot actions.

V. Experiment Results

To evaluate the complete system, we used the data from the *cooking assistance* dataset. The input of our system is the complete speech of the participant for a task and the output of our system is the graph-based sequence of actions. Our testing set consists of the data from the five participants who were excluded from the training dataset. The accuracy, which is the number of correctly generated robot actions divided by the total number of robot actions requested by the participant, is then calculated for the testing set. The accuracy of the system is calculated to be **98.6%**. Moreover, the BiLingual Evaluation Understudy (BLEU score) is also calculated, which ranges from 0 to 1, with 0 meaning no match between the generated and target sequence of actions to 1 being an exact match between the generated and target sequence of actions. The BLEU score for the system is **0.9847**.

The results are very positive and there are very few cases in which our system fails; however, additional data from more participants may be needed to verify the results in a larger testing sample. It is also important to discuss the failures of our system. For example, the sentence "Hey Robot, can you give me some more?" is classified as a valid command and added to the graph. This is not correct as it is not clear what the robot should actually fetch more of. This case is more of a limitation than a false prediction because when the participant was giving the command to the robot, there was a long pause between "Hey Robot, can you give me some more" and "water". This is the limitation of the system where the user cannot give a long pause in between giving a command to the robot. Another example that our system failed was because it replaced some words with synonyms; for example, the participant asked for a pot but the generated robot action was "bring pan". The words are very near in meaning but nevertheless different.

VI. Conclusion and Future Work

In the presented work, a collaborative cooking scenario between a human and a robot is introduced. The human communicates with the robot verbally and provides instructions on how it can be of assistance. We propose a framework that automatically generates robot actions from speech that included environmental interruptions. A small study of 15 people was conducted to evaluate our system. The accuracy of the system is 98.6%, which is a very promising and positive finding. However, the system has a few failed cases.

In the future, we plan to have a more extensive study to ensure our proposed framework is robust and the dataset will be published. Additionally, the graph-based sequence of actions will be connected to a robot skill library, similar to [25], that would enable the robot to perform the actions in real time.

Acknowledgment

This material is based upon work supported partially supported by the National Science Foundation under Grant 2226165 and by Santa Clara University under University Research Grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] L. Grasse, S. J. Boutros, and M. S. Tata, "Speech interaction to control a hands-free delivery robot for high-risk health care scenarios," *Frontiers in Robotics and AI*, vol. 8, 4 2021.

[2] T. Kumar Shastha, M. Kyrarini, and A. Gräser, "Application of reinforcement learning to a robotic drinking assistant," *Robotics*, vol. 9, no. 1, p. 1, 2019.

[3] F. F. Goldau, T. K. Shastha, M. Kyrarini, and A. Graser, "Autonomous multi-sensory robotic assistant for a drinking task," *IEEE International Conference on Rehabilitation Robotics*, vol. 2019-June, pp. 210–216, 6 2019.

[4] M. Vila Abad, G. Canal, and G. Alenyà, "Towards safety in physically assistive robots: eating assistance," 2018.

[5] S. Gushi, H. Higa, H. Uehara, and T. Soken, "An assistive robotic arm for people with severe disabilities: evaluation of eating soup," *J. Advanced Control, Automation and Robotics*.

[6] D. E. Gerber, "Eatdis research project," 2017, [Accessed: Oct 12, 2022]. [Online]. Available: https://msuweb.montclair.edu/~gerbere/food-disability-research-2017.1.html

[7] Y. Sugiura, D. Sakamoto, A. Withana, M. Inami, and T. Igarashi, "Cooking with robots: designing a household system working in open environments," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2010, pp. 2427–2430.

[8] M. Marge, C. Espy-Wilson, N. G. Ward, A. Alwan, Y. Artzi, M. Bansal, G. Blankenship, J. Chai, H. Daumé III, D. Dey *et al.*, "Spoken language interaction with robots: Recommendations for future research," *Computer Speech & Language*, vol. 71, p. 101255, 2022.

[9] D. Strazdas, J. Hintz, A.-M. Felßberg, and A. Al-Hamadi, "Robots and wizards: An investigation into natural human–robot interaction," *IEEE Access*, vol. 8, pp. 207 635–207 642, 2020.

[10] V. V. Unhelkar, S. Li, and J. A. Shah, "Decision-making for bidirectional communication in sequential human-robot collaborative tasks," *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. [Online]. Available: https://doi.org/10.1145/3319502.3374779

[11] A. González-Docasal, C. Aceta, H. Arzelus, A. Álvarez, I. Fernández, and J. Kildal, "Towards a natural human-robot interaction in an industrial environment," *Lecture Notes in Electrical Engineering*, vol. 704, pp. 243–255, 2021.

[12] D. Choi, W. Shi, Y. S. Liang, K. H. Yeo, and J. J. Kim, "Controlling industrial robots with high-level verbal commands," *Lecture Notes in Computer Science including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, vol. 13086 LNAI, pp. 216–226, 2021.

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2018. [Online]. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[14] C. Li, X. Zhang, D. Chrysostomou, and H. Yang, "Tod4ir: A humanised task-oriented dialogue system for industrial robots," *IEEE Access*, pp. 1–1, 8 2022.

[15] Institutional Review Board (IRB) and Protection of Human Subjects in Clinical Trials. [Accessed: Oct 12, 2022]. [Online]. Available: https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/institutional-review-boards-irbs-and-protection-human-subjects-clinical-trials

[16] "Rosbags wiki," [Accessed: Oct 12, 2022]. [Online]. Available: http://wiki.ros.org/Bags

[17] C. Deuerlein, M. Langer, J. Seßner, P. Heß, and J. Franke, "Human-robot-interaction using cloud-based speech recognition systems," *Procedia CIRP*, vol. 97, pp. 130–135, 1 2021.

[18] "Vosk models," [Accessed: Oct 12, 2022]. [Online]. Available: https://alphacephei.com/vosk/models

[19] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North*, pp. 4171–4186, 2019. [Online]. Available: https://aclanthology.org/N19-1423

[20] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *arXiv preprint arXiv:1506.06724*, 2015.

[21] FoodNetwork.com, "Recipes a to z," 2022, [Accessed: Oct 12, 2022]. [Online]. Available: https://www.foodnetwork.com/recipes/recipes-a-z/123

[22] W. Xiong, J. Wu, H. Wang, V. Kulkarni, M. Yu, S. Chang, X. Guo, and W. Y. Wang, "Tweetqa: A social media focused question answering dataset," *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 5020–5031, 2019. [Online]. Available: https://aclanthology.org/P19-1496

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[24] "Huggingface distilgpt2," [Accessed: Oct 12, 2022]. [Online]. Available: https://huggingface.co/distilgpt2

[25] M. Kyrarini, S. Naeem, X. Wang, and A. Gräser, "Skill robot library: Intelligent path planning framework for object manipulation," *25th European Signal Processing Conference, EUSIPCO 2017*, vol. 2017-January, pp. 2398–2402, 10 2017.