

Enhancing Interaction in Industrial Collaborative Robots with Advanced AI Solutions

Ignazio Infantino^{1,*†}, Carmelo Mineo¹

¹*Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR), Consiglio Nazionale delle Ricerche (CNR), Via Ugo La Malfa 153, Palermo, Italy*

Abstract

Collaborative robots allow humans to act safely in industrial environments, and current artificial intelligence capabilities enhance their interaction using multi-modal approaches: speech and vision based on neural deep networks (DNNs) and natural language processing based on Large Language Models (LLMs) drastically improve the potentiality to perform complex collaborative tasks. The paper shows how to combine them in real architecture in a complex product assembling scenario. Moreover, it reports a short discussion about the future direction that could enable an effective and efficient use of modern AI approaches in the Industry 5.0 framework.

Keywords

Collaborative Robot, Human-robot Interaction, Artificial Intelligence, Cognitive Architectures, Digital Twin

1. Introduction

In the context of Industry 5.0 [1], the new generation of collaborative robots (or CoBot) have to act in the presence of human operators safely and have the role of effective work companions. That implies several aspects to consider to establish a fruitful and satisfactory work relationship [2]. The future vision points to “*human-robot co-working*”, where humans and robots collaborate closely on various complex tasks. In this scenario, humans would concentrate on tasks that require creativity, innovation, and complex problem-solving, while robots handle repetitive, labor-intensive, or precision-driven tasks. This partnership aims to optimize efficiency and productivity by leveraging the strengths of both human ingenuity and robotic precision. However, implementing human-robot co-working brings several challenges that go beyond technical innovations. One significant issue is the evolution of organizational behavior and structure. As companies adopt this collaborative approach, workflows must be restructured to accommodate human and robotic contributions. Moreover, businesses must consider the impact on the work environment, from physical layout changes that allow robots to operate effectively alongside humans to adjustments in team dynamics and leadership models. Privacy and trust are key concerns, particularly when humans and robots share sensitive information or work closely together. Establishing trust between humans and their robotic counterparts and ensuring data security in these interactions will be vital for the success of this vision. In the paper, we propose the implementation and some design refinements of the robotic system presented in [3]. Speech understanding, artificial vision, and verbal interaction allow the system to interact with the human in a natural way to pursue several goals: to monitor a complex manipulative assembling task, to mutually exchange information on the sequential steps to complete and involved components, to share with human the reasoning process underlying the decisions and the actions of the robotic companion. The proposed system is coupled with its digital twin that owns the same interaction capabilities as the natural robotic system and allows the operator to simulate collaborative task execution realistically.

11th Italian Workshop on Artificial Intelligence and Robotics (AIRO 2024)

*Corresponding author.

†These authors contributed equally.

✉ ignazio.infantino@cnr.it (I. Infantino); carmel.mineo@cnr.it (C. Mineo)

id 0000-0002-0980-7391 (I. Infantino); 0000-0002-5086-366X (C. Mineo)



© 2025 This work is licensed under a “CC BY 4.0” license.

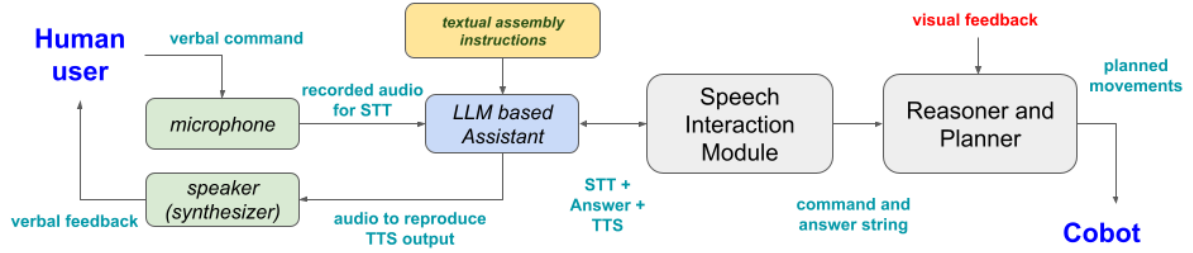


Figure 1: The main components of the software cognitive architecture responsible for processing sensory inputs and deciding and performing actions

2. Developing CoBot’s interaction capabilities

Large Language Models (LLMs) and, in general, Deep Neural Networks (DNNs) allow artificial intelligence systems to show complex interaction capabilities. Particular software architectures known as Cognitive Architectures (CAs) can reproduce human sensing [4] and reasoning processes [5], knowledge management, social interaction [6], and planning. They enable robots to act as humans for many behavioral aspects (expectations, personality, creativity, and emotions) [7]. Such an approach supports establishing a natural interaction between humans and robots to ensure desirable features such as trustability, accountability, transparency, and explainability.

At present, we provide the system of a simple implementation of a cognitive architecture that includes basic sensing capabilities (vision and speech), short-term memory, and a planner/decision module to execute all sequential steps of the assembly task. The aim is to have a verbally interactive prototype that allows us to experiment with the interaction with a human in a real environment and to accomplish a complex collaborative task. In future work, after collecting several tests of such interactions both in simulated and real environments, we aim to expand the architecture with higher cognitive capabilities and more sensory inputs such as, for example, the touch. Figure 1 shows the main components of the software cognitive architecture responsible for processing sensory inputs and deciding and performing actions. In the following, we describe AI-based functionalities to recognize speech, process visual streams from a camera placed on the cobot actuator, and interact verbally.

2.1. Speech understanding

Verbal interaction is the most simple and direct way humans use daily. If a robotic agent has verbal capabilities [8], the human operator needs to have only the necessary knowledge for the specific task execution. To process speech, the system needs different software components that are responsible for the following:

- Speech to text (STT) conversion
- Engagement
- Dialogue management (DM)
- Command recognition (e.g. pick x, find y, next x, info task, stop\abort\reset, ...)

By engagement, we mean a possible initial phase preceding the execution of the task, which may involve the operator’s recognition or otherwise and the indication of the possible verbal interaction functionalities.

The complexity of verbal interaction between humans and robots requires a high precision rate of recognition and real-time performance. Moreover, developing interactive systems that assure privacy and security in the industrial environment could be of primary interest. Accessing remote cloud services (such as those provided by most relevant enterprises) ensures top performance in recognizing speech and finding suitable answers both in specific and general knowledge domains. Suppose we use only local computation resources to manage privacy. In that case, security issues and the protection of the

enterprise's valuable data require high financial investments to build personalized solutions. That also requires continuous updating to have better performance and robustness. Open-source approaches allow cheaper solutions with acceptable performances, especially if the verbal interaction is limited to a more specific context and domain. In particular, the system uses pre-trained language models (both in Italian and English language) of the speech recognition toolkit Vosk [9]. The robust STT module allows the system to process audio input from a wireless microphone and give textual input to recognize commands with parameters or general requests. Verbal commands cause the robotic arm to execute specific actions, while other requests cause verbal feedback by an LLM-based assistant. We locally use a pre-trained Vosk model to ensure maximum privacy. Recognition of the command is linked to a confidence threshold: if it is below the limit or there is an empty string, you are asked to repeat the verbal command. A database stores the set of main commands linked to executing actions.

2.2. Vision

The vision module processes frames acquired by the RealSense depth camera D435i [10] to detect and recognize relevant objects to support the understanding of assembly task evolution. A deep learning approach allows the system to classify with precision and accuracy all objects handled by the human and the cobot, as well as occlusions and critical light conditions. Starting from YOLO v2 model [11], we include new classes of objects by single view and using data augmentation. We also test a more complex object detection algorithm processing depth and RGB images to perform action prediction proposed in [12]. Moreover, in the case of a fixed set of objects, the system can use classical feature-based approaches (e.g. SIFT) to reduce communication and computational costs using simple grayscale frames. The vision module also estimates object position and orientation in the reference system. Object recognition allows to monitor the assembly task by verifying the adherence to the mounting sequence and the availability of all necessary objects in the storing area to accomplish the next steps.

2.3. LLM

The human operator can request various information from an LLM-based assistant by accessing remote cloud services. The assistant knows the textual assembling instructions, the list of objects, and their descriptions (colour, shape, position in storage areas, number of pieces). In this way, the cobot can instruct the human operator how to accomplish a given assembly task by describing the sequence of different phases and actions to do (start, stop, continue, repeat, pick, place, rotate, and so on). The collaborative task could evolve in two ways: Case 1 - The human operator leads the task. The robot continuously observes the scene and gives the user verbal feedback on the task's evolution. The user can ask the robot to perform some actions. Case 2 - The Robot leads the assembly task. The robot conducts the task by explaining its actions and state. It asks the human to perform difficult or impossible actions, such as when there are unreachable or missing components or the gripper tool is not suitable for some object picking or manipulation because of fragility or deformability. The verbal assistant could provide details about the performances of the various processing modules, the current state and possible evolution of robot behavior, and its hypothesis on human expectations.

3. Exploiting a real scenario: assembly task

The chosen application test case is based on a MECCANO toy kit. The kit contains all the components required to assemble a toy model of a motorbike (see Figure 2a). It is composed of 49 components with different shapes and sizes. Similarly to what happens in an industrial scenario, the subjects interact with tools such as a screwdriver and a wrench, as well as with tiny objects such as screws and bolts while executing a task involving sequential actions (e.g., take the wrench, tighten the bolt, put down wrench). Even though this scenario simplifies what can be found in real-world industrial settings, it is reasonably complex. Moreover, this is the same toy kit used by the authors of [12], where a benchmark to study human behavior in an industrial-like scenario is proposed. Furthermore, the datasets generated

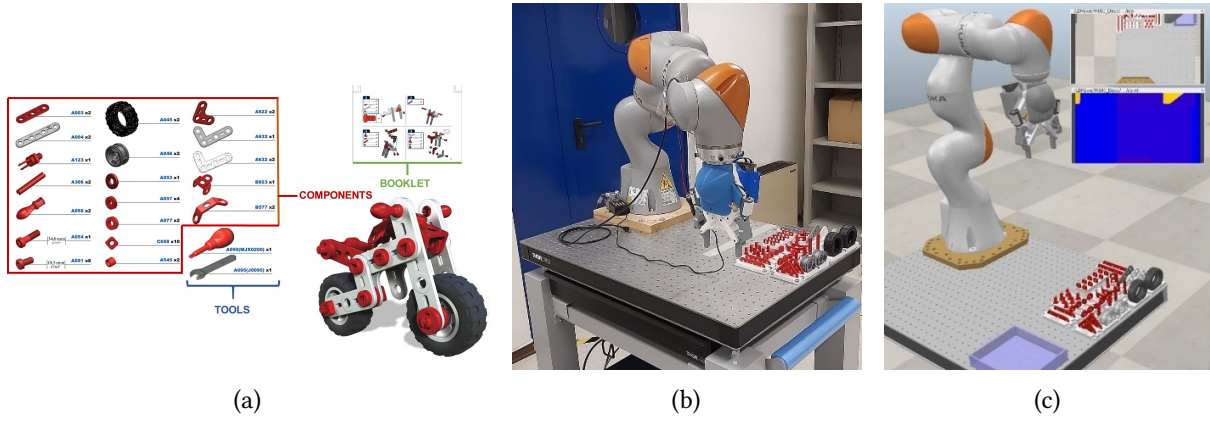


Figure 2: MECCANO toy kit (a), real-world setup (b), and digital setup (c).

in [12] are publicly available and may be used in the upcoming phases of this ongoing work. The robotic system used here is based on a KUKA iiwa 7 R800 collaborative robot (see Figure 2b) and an in-house developed robotic gripper (consisting of a LEGO NXT Mindstorms controller, motors and gears [13] and 3D-printed body parts). Figure 2c shows the digital and real-world representation of the setup.

3.1. Examples of verbal interaction

Verbal commands such *PICK A003 2* (i.e. pick two pieces of component A003) cause the robot arm to perform all the possible actions in real and virtual environments. After the execution is completed, the system waits for further instructions (verbally or by user interface). Naturally, the planner monitors the right sequence execution and eventually gives notice by a synthesized voice. In the following, we also report some examples of verbal interactions that allow the human operator to have generic information on the task.

Human operator: What can I build with the kit? *Assembly Assistant:* The Meccano Junior 16102 kit allows you to build 3 different motorcycles. Each motorcycle is built by following a sequence of assembly steps. Here's an overview of the components included in the kit and their quantities: Red bar with three holes (A003) - 2 pieces; Gray bar with 4 holes (A004) - 2 pieces; ...

Human operator: How do I start? *Assembly Assistant:* To start building one of the motorcycles with the Meccano Junior 16102 kit, we can follow the assembly sequence indicated in the manual. Phase 1 of the assembly uses the following components: L-shaped white bars with 4 holes (A632-b); Short red bolts (A051); Gray bars with 4 holes (A004); ...

Human operator: Which pieces are the most numerous? *Assembly Assistant:* Among the pieces included in the Meccano Junior 16102 kit, the most numerous are: Red square washers (C658) - 10 pieces; Short red bolts (A051) - 8 pieces; ...

Human operator: What types of bars are there? *Assembly Assistant:* In the Meccano Junior 16102 kit, there are different types of bars used to build the motorcycles. Here's a list of the various types of bars included in the kit: Red bar with three holes (A003) - 2 pieces; Gray bar with 4 holes (A004) - 2 pieces; L-shaped red bar with three holes (A622) - 2 pieces; ...

Human operator: What actions are required in Phase 1? *Assembly Assistant:* In the first phase of assembling the first motorcycle with the Meccano Junior 16102 kit, the actions to take are as follows: Insert two long screws (code A631) into the base's holes (code A006); Mount the two tires (code A051) on the short wheel axes (code A632-a); ...

3.2. Digital twin

This work has created a digital twin of the system (Figure 2c), whose purpose is twofold: it allows offline simulation and online monitoring of the assembly process. CoppeliaSim software was used

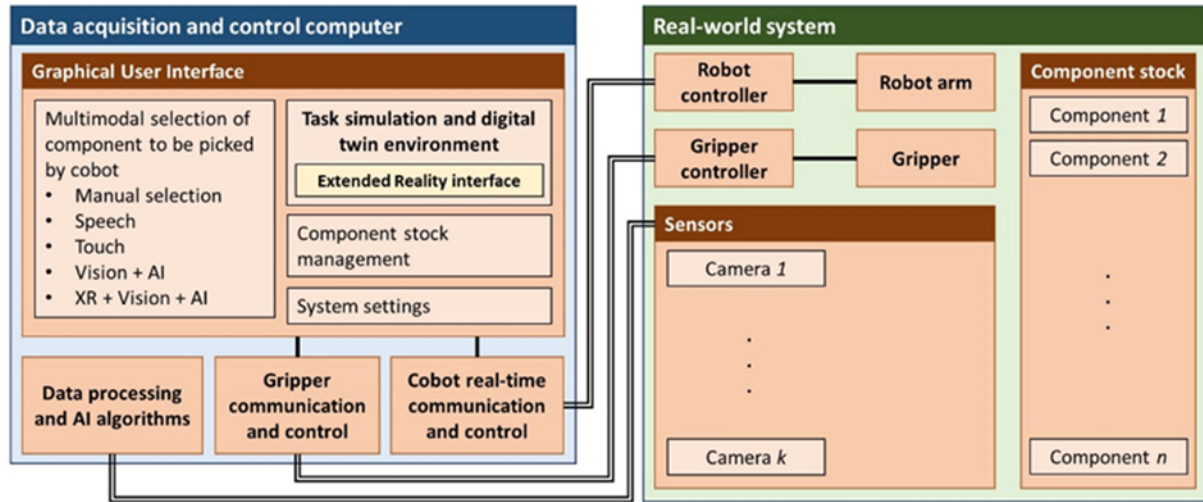


Figure 3: Software architecture for human-robot interaction.

to create the digital environment. This software uses algorithms dedicated to calculating direct and inverse kinematics and different simulation modes of the physical world (MuJoCo, Bullet, ODE, Vortex, Newton Game Dynamics) for the motion of rigid bodies. The digital models and scenes are built by assembling various objects (meshes, joints, sensors, point clouds, etc.) in a hierarchical and relational structure. The capability of CoppeliaSim to run multiple software threads simultaneously has been exploited to link the digital system representation with a purposely developed graphical user interface (developed in Matlab) and mirror the feedback data received from the physical world. The use of the Kuka Sunrise Toolbox (KST), presented in [14], has enabled real-time control of the cobot. The custom gripper is controlled through a Dynamic Link Library (DLL), programmed and compiled through the LabView programming environment. This DLL enables communication with the gripper's LEGO NXT Mindstorms control unit. When the digital twin is coupled with the physical robotic system, both systems perform the procedures in real-time, which could allow remote control. The digital twin is also designed as a simulator to experiment and optimize assembly tasks without using the robotic arm.

4. Discussion and future work

The proposed system aims to investigate human-robot collaboration further from different points of view. If we consider a dynamic execution and the possibility of varying the action plan during the work session, humans and robots have to collaborate at different levels: physically (avoiding dangerous collisions, unnecessary movements, or detecting human touch), action level (performing a shared action plane to accomplish the task), cognitive level (exchanging knowledge and experiences, understanding mutual intentions, expectations, satisfaction levels, physical wellness. The robotic system should be *aware* of the skill and the behavior of the human companion by taking into account different aspects:

- execution velocity
- execution times
- estimation of performances: error rate, fatigue, distraction, ...
- interaction level: talkativeness, degree of verbal interaction, willingness to argue

Moreover, suppose the robot has a high level of autonomy. In that case, we also have to consider moral and ethical issues when critical decisions have to be made by machines (or artificial agents).

- How do we evaluate human-robot interaction?
- metrics on task execution (time, precision)
- degree of human satisfaction (collaborator, remote operator)

- the instauration of a relationship (as a companion)
- empathy and emotional involvement
- trustability and accountability (based also on self-explainability)

It occurs to define also the (social) role of the robot concerning the human companion: same functionalities (interchangeable during execution); complimentary, i.e., each one with a specific assignment; master (robot as coordinator) or enslaved (task execution driven by humans)

Acknowledgments

This work has been funded by the "Progetto Integrato Tecnologie dell'Idrogeno" (CUPB53C22008610001), part of the Three-Year Implementation Plan 2022-2024 for the Italian National Electricity System Research (Research Topic 1.3).

We want to thank the Competence Center ARTES 4.0 (Advanced Robotics and Enabling Digital Technologies and Systems) for their contribution to the hardware implementation of the collaborative robotics demonstration line.

References

- [1] K. A. Demir, G. Döven, B. Sezen, Industry 5.0 and human-robot co-working, *Procedia Computer Science* 158 (2019) 688–695. doi:<https://doi.org/10.1016/j.procs.2019.09.104>, 3rd World Conference on Technology, Innovation and Entrepreneurship, June 21-23, 2019.
- [2] A. Adel, Future of industry 5.0 in society: human-centric solutions, challenges and prospective research areas, *Journal of Cloud Computing* 11 (2022). doi:10.1186/s13677-022-00314-5.
- [3] C. Mineo, I. Infantino, O. Barbera, G. Giaccoppo, N. Montinaro, Sustainable by design: Advancing hydrogen fuel cell production through automation, *MATEC Web Conf.* 401 (2024) 12001. URL: <https://doi.org/10.1051/mateconf/202440112001>. doi:10.1051/mateconf/202440112001.
- [4] A. Chella, H. Dindo, I. Infantino, A cognitive framework for imitation learning, *Robotics and Autonomous Systems* 54 (2006) 403–408.
- [5] I. Infantino, G. Pilato, R. Rizzo, F. Vella, Humanoid introspection: A practical approach, *International Journal of Advanced Robotic Systems* 10 (2013) 246.
- [6] I. Infantino, A. Augello, U. Maniscalco, G. Pilato, F. Vella, A cognitive architecture for social robots, 2018 IEEE 4th International Forum on Research and Technology for Society and Industry (RTSI) (2018) 1–5.
- [7] I. Infantino, Affective human-humanoid interaction through cognitive architecture, *The Future of Humanoid Robots-Research and Applications* (2012) 147–164.
- [8] M. Marge, C. Espy-Wilson, N. G. Ward, Al., Spoken language interaction with robots: Recommendations for future research, *Computer Speech Language* 71 (2022) 101255. doi:<https://doi.org/10.1016/j.csl.2021.101255>.
- [9] Vosk speech recognition toolkit (2024). URL: <https://alphacephei.com/vosk/>, accessed 2024-09-30.
- [10] P. Schmidt, J. Scaife, M. Harville, S. Liman, A. Ahmed, Intel® realsense™ tracking camera t265 and intel® realsense™ depth camera d435-tracking and depth, *Real Sense* (2019).
- [11] Ultralytics yolo v2 model (2024). URL: <https://docs.ultralytics.com/datasets/detect/coco/>, accessed 2024-09-30.
- [12] F. Ragusa, A. Furnari, G. M. Farinella, Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain, *Computer Vision and Image Understanding* 235 (2023) 103764.
- [13] J. Gonçalves, J. Lima, P. Costa, Rapid prototyping of mobile robots extending lego mindstorms platform, *IFAC Proceedings Volumes* 39 (2006) 47–52.
- [14] M. Safeea, P. Neto, Kuka sunrise toolbox: Interfacing collaborative robots with matlab, *IEEE Robotics & Automation Magazine* 26 (2018) 91–96.