Co-speech Gestures for Human-Robot Collaboration

Akif Ekrekli
Engineering and Natural Sciences
Tampere University
Tampere, Finland
akif.ekrekli@tuni.fi

Alexandre Angleraud
Engineering and Natural Sciences
Tampere University
Tampere, Finland
alexandre.angleraud@tuni.fi

Gaurang Sharma
Engineering and Natural Sciences
Tampere University
Tampere, Finland
gaurang.sharma@tuni.fi

Roel Pieters
Engineering and Natural Sciences
Tampere University
Tampere, Finland
roel.pieters@tuni.fi

***Abstract*—Collaboration between human and robot requires effective modes of communication to assign robot tasks and coordinate activities. As communication can utilize different modalities, a multi-modal approach can be more expressive than single modal models alone. In this work we propose a co-speech gesture model that can assign robot tasks for human-robot collaboration. Human gestures and speech, detected by computer vision and speech recognition, can thus refer to objects in the scene and apply robot actions to them. We present an experimental evaluation of the multi-modal co-speech model with a real-world industrial use case. Results demonstrate that multi-modal communication is easy to achieve and can provide benefits for collaboration with respect to single modal tools.**

***Index Terms*—Human-robot collaboration, multi-modal perception, speech recognition, gesture detection, object detection**

I. Introduction

Fluent interaction between human and robot requires reliable perception to capture the commands of a person. While recent approaches in deep learning [1] have established impressive tools to detect e.g., human pose, gestures and speech, single tools alone can not always convey easily the commands intended [2]. Reasons for this are the limited expressions available for different modes of communication and the limitations in perception performance. Human hand gestures, for example, contain much less information content than speech. On the other hand, gesture detection can be done much quicker than speech recognition, leading to a faster response time. These conflicting properties motivate to combine multiple perception tools into a single multi-modal detection model that utilizes communication from human to robot for assigning tasks and coordinating the collaboration. In this work we compare different perception tools and analyse them with respect to their suitability for human-robot collaboration. A co-speech gesture model is then developed that combines speech, human hand gestures and object detection to achieve effective communication of desired robot tasks, such as picking human-specified objects and robot to human hand-overs (see Fig. 1). The developments are intended for industrial human-robot collaboration where a collaborative robot shares its tasks, and works in close collaboration with, a human operator. Our contributions are:

- Human speech and hand gesture perception methods to command robot actions
- Co-speech gesture model that combines human natural speech and hand gestures to command robot actions
- Experimental evaluation of the co-speech gesture model in an industrial human-robot collaborative use case

Project funding was received from EU's Horizon 2020 research and innovation programme, grant no. 871449 (OpenDR).

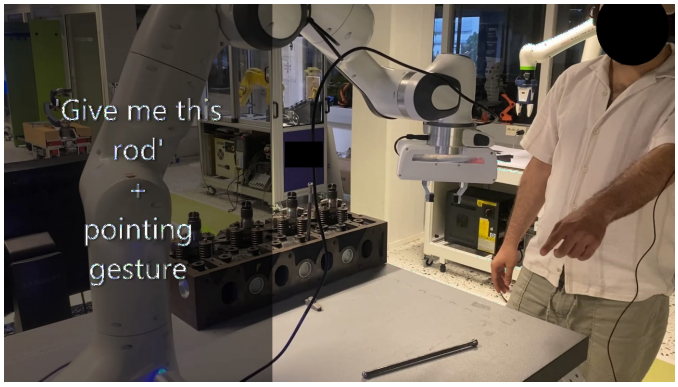


Fig. 1: Co-speech gesture model that combines a speech phrase, human gesture detection and object perception to command robot actions.

II. Related work

A. Human-Robot Collaboration

Collaboration between human and robot is often targeted for industrial manufacturing [3], as both robot and human have unique skills that complement each other. Different interfaces that enable the collaboration have been analyzed, providing clear directions on how the collaboration benefits the tasks [4]. Approaches include voice processing, gesture recognition, haptic interaction, and even brainwave perception. Often machine [5] and deep [6] learning are used as enabling perception tool [1] to classify and recognize the person and objects in the environment [7].

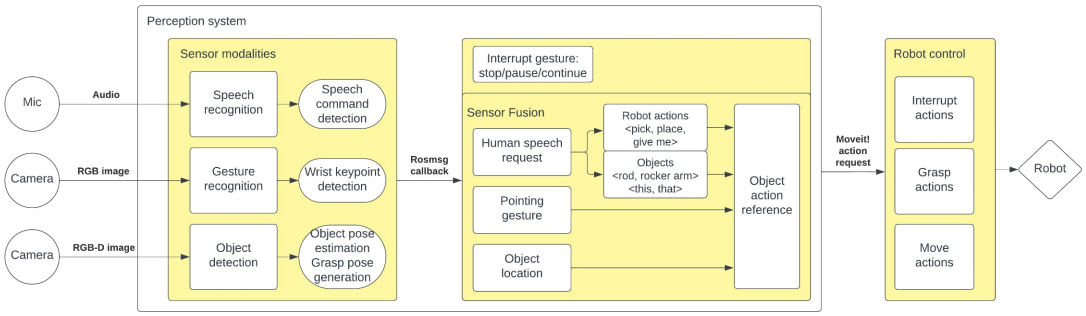


Fig. 2: Co-speech gesture model that takes input from speech commands, gesture recognition and object detection to generate robot actions for human-robot collaboration. Sensor fusion enables the human to refer to specific objects (<rod, rocker, arm, this, that>) and apply actions to them (<pick, place, give me>).

B. Human Perception

Visual detection of a person in the scene has been an active area of research [8]. Different visual modalities have been utilized [9], such as RGB and depth information [10]. Multi-modal approaches that utilize RGB-D data are popular as well [11]. Human pose estimation goes a step further than human detection by estimating the 3D pose of a human and their individual skeleton joints [12], which can be used as input for gesture detection. Utilizing speech for commanding robots has been demonstrated with short verbal commands for task coordination [13] and task programming [14]. As extension to short speech commands, natural language as instructions to robots has been used for planning [15] and allocation [16] of tasks to be performed by the robot.

Multi-modal human-robot collaboration using gestures and speech simultaneously has been demonstrated for a human interacting with the humanoid robot NAO in [17], where short phrases and gestures are utilized to indicate human actions. Collaboration between a robot arm and a human worker is also demonstrated in [18], where a set of gestures and speech commands are perceived individually to produce the same input for robot actions. As comparison, our work considers an industrial scenario with a collaborative robot where speech phrases and gestures are combined to assign tasks to the robot.

III. Methods and Tools

A. Perception Tools

The perception tools utilized in this work are integrated in a common framework for isolated and human-robot collaborative tasks. For human perception, Lightweight OpenPose, a human skeleton detection tool [12] is used, which takes images (RGB) as input and returns skeleton node points as output. For interaction, the wrist node of the skeleton is taken and, when presented in a certain image area, serves as trigger for robot actions (e.g., stop, continue) or refers to certain objects in the scene (i.e., detected objects pointed to). In the latter case, the detected object that is closest to the wrist node is selected for robot action execution. Speech recognition is enabled by Vosk [19] for the detection of pre-defined input commands and phrases. This set of words and sentences relate to available actions of the robot and locations in the scene, as described in Table I. The model is configured by filtering out unnecessary words that are unsuitable for robot instructions. Objects in the scene are detected by a neural network (Detectron2 [20]) trained on a custom dataset collected for the use case [21].

B. Multi-modal Perception Methods

The perception tools can be used in different ways to allow for sensor redundancy, sensor multi-modality and sensor information fusion, as follows.

- **Sensor redundancy** - multiple sensors are used to command the same robot actions, e.g., speech or hand gesture to stop robot motion
- **Sensor multi-modality** - different sensor modalities are used to command individual robot actions, e.g., speech provides the robot actions, vision detects human gestures
- **Sensor-fusion** - different sensor modalities are combined to command a single robot action, e.g., speech provides robot action, vision provides specific object location as pointed to by the human

While sensor redundancy and multi-modality is supported and demonstrated in Section IV, we emphasize our contributions to the fusion of multiple sensor outputs into a single robot command, as explained in the following section.

C. Co-speech Gesture Model

The single-modal visual and speech perception models are fused into a multi-modal perception model by combining speech commands, pointing gestures and object detection (see Fig. 2). Several examples of these co-speech gestures are described in Table I. The human can refer to individual objects in the scene by speech (e.g., <rod>, <rocker arm>) and pointing to them, and apply specific robot actions by speech commands (e.g., picking with <pick>, placing with <place>, robot to human hand-over with <give>).

Depending on the object, different robot actions are possible, as specified beforehand. For example, objects can be picked up from the table, placed in specified locations and handed over to the person. Object detection returns a list

TABLE I: Perception methods' input and output

Method	Input	Output
Wrist detection	RGB image of the scene (human front-facing) Human gesture by moving wrist to certain image location	Robot stop/continue actions
Speech recognition	Robot action commands: <pick, place, give, go, stop, pause, continue> Workspace commands: <rod, home, arm, me> Human speech requests: <place rod>, <go home>, <give me another rocker arm>, <pick up the last rod>	Robot motion Gripper actions Robot to human hand-over Robot stop/continue actions
Object detection	RGB image of the scene (top-down)	Detected objects in the scene Valid target location for robot
Co-speech gesture	<pick rod> + pointing gesture + object detection <give me this rod> + pointing gesture + object detection <give me that rocker arm> + pointing gesture + object detection	Robot motion Gripper actions Robot to human hand-over

of objects in the scene, which can be verbally referred to by their class. Pointing gesture detection allows to refer to specific objects in the scene by relating the pointing gesture location to detected object locations. Robot actions are therefore commanded by specific action verbs and object classes, complimented by gestures to provide fine-grained object references (see Fig. 2).

IV. Experimental Results

A. Industrial Use Case

The considered use case replicates an industrial assembly task that in current situation is done manually by human operators. The solution we propose introduces a collaborative robot as assistive tool to the assembly station, under control of the person. This means that the assembly work is coordinated by the human, with the robot assisting in tasks that the human decides. Available robot actions are to move to certain locations in the work space, pick objects that are detected on the table, place objects to specified locations or hand them over to the human. In addition, coordinated actions include the stopping and continuing of robot actions during execution, for human visual inspection of the objects placed by the robot. Human commands can be communicated by hand gestures and/or speech, with different levels of functionality as described in Table I. The setup for experiments is depicted in Fig. 3 and includes two cameras (Intel Realsense D435) for visual perception (one front-facing for wrist detection; Fig. 3(b) and one top-down for object detection; Fig. 3(c)) and a microphone for speech recognition. Computation is performed on a standard Desktop PC running Ubuntu Linux with Nvidia GTX 1080 Ti GPU, and all robot (Franka Emika) communication and control utilizes ROS. All tools are open-source available to utilize or replicate: https://github.com/opendr-eu/opendr.

B. Human Gesture Detection

Results for the visual wrist detection tool are depicted in Fig. 3(b), which highlights both detected human wrists. When one of the wrists is detected inside one of the squares, this is taken as trigger for referring to certain robot actions or objects in the scene. For example, to stop robot motion, the left wrist should be detected in the top left square and to continue robot motion, the right wrist should be detected in the top right square. Pointing gestures are interpreted in a similar manner. When the human points to a certain object, first the left or right wrist needs to be detected in either of the lower two squares in the image, after which the location to the closest detected object is determined. Performance of the skeleton detection tool has been reported in the original paper [12]. In our use case the detection accuracy of the wrists inside a square is consistent around 90%, as assessed from 20-second interval tests for different squares. This is satisfactory for effective collaboration.

C. Object Detection

Results of visual object detection are depicted in Fig. 3(c), which has the different detected objects annotated by colored bounding boxes (yellow for the rocker arms and blue for the rods). As objects are detected in image space, careful calibration of both cameras ensures the detected objects can be picked from the table and that pointing gestures can refer to the same object in both camera frames. In our use case the detection accuracy of all classes is over 90%.

D. Speech Recognition

Results of speech recognition were found satisfactory, as in most cases the spoken commands are recognized correctly. Performance, as reported in [19], depends on the language skills of the person giving commands, as in certain cases non-native English speakers had to speak more clear to achieve correct speech recognition. Besides the speech recognition itself, the speech tool was improved by including a voice activity detector and a time-delay filter (0.5 seconds) to consider the natural pause in human speech. This resulted in a delay of ≈ 1.9 seconds between a verbal command and the recognized speech (average of 50 trials with different commands).

E. Co-speech Gesture Model Performance

The co-speech gesture model has all three perception models running in parallel, decreasing slightly the running performance of the skeleton detection tool (i.e., 24 fps with image size of 1920×1080). Object detection achieves a frame rate of 4.5 fps with image size of 1280×720. Extended experiments were performed to test the co-speech tool in a collaborative assembly scenario. This included a human and robot performing assembly steps to an engine, with parts that are either mounted by the person or by the robot. Parts assembled by the person are picked by the robot from the

(a) (b) (c)

Fig. 3: Experimental setup with a human pointing at an object for robot picking (a). One camera is human front-facing to capture human hand gestures (b), one camera is mounted on the robot (eye-in-hand) for object detection on the table (c).

table and handed over to the human, and parts assembled by the robot are picked by the robot from the table and directly mounted to the engine. Coordination of the tasks and requesting robot actions is done by the person via the co-speech gesture model. In addition, the human can halt and continue robot tasks at any time, by both gesture and speech commands (<stop>, <pause>, <continue>).

Single commands **-** Fig. 4(a-b) depict the human commanding a stop and continue gesture, respectively. Fig. 4(c) shows the human commanding the robot to move to its 'home' configuration by the phrase <ok, go home>. For this, the home location is preprogrammed in the software scripts.

Speech phrases **-** Fig. 5 depicts how human speech alone can be utilized to command robot actions, by the phrase <give me another rocker arm>. From the recognized speech, the tool extracts relevant words and connects these to robot actions and objects in the scene. In this case <give me> refers to a robot to human hand-over, <rocker arm> refers to the rocker arm class in the object detection model, and <another> implies any of the detected rocker arms, meaning the first in the returned detection list. As a result, the command phrase initiates all required robot actions and starts executing them one-by-one, as shown in Fig. 5(ac).

Co-speech commands **-** Fig. 6 depicts examples of the co-speech gesture model that utilizes a human speech phrase and pointing gesture to achieve robot actions applied to specified objects in the scene. In this case, as a pointing gesture is detected by the wrist detection tool, the closest specified object to the human wrist is selected for the robot actions. A video of the co-speech gesture model demonstrates all commands from Fig. 4-6: https://youtu.be/b_ISrhOlcC8. This shows the collaborative tasks, where the human coordinates the actions of the robot with four pick and place actions and four robot to human hand-overs. Human inspection is done after object placement by stopping robot motion with a speech command. In total, the experiment includes over 20 speech commands and seven co-speech gestures to coordinate the shared task.

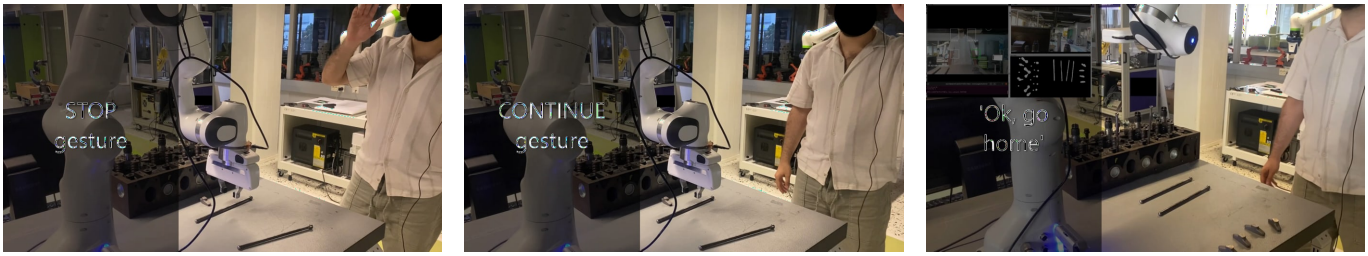
V. Discussion and Limitations

Sensor redundancy enables different modalities to command the same robot action. This was demonstrated for stopping and continuing robot motion and actions by hand gestures (see Fig. 4) and by speech commands. While hand gestures can be detected at relatively high rate (>24 FPS), it can takes several image frames before a correct prediction occurs. On the other hand, speech commands can have considerable delay even when a first verbal command is correctly recognized.

While in most cases the co-speech gesture model achieves the intended robot commands and collaboration, some limitations are identified. First, detection of the human wrist in a specific image location requires careful human hand motion. As alternative, human hand gestures could be recognized directly from a dedicated model [22]. In our case, inference time and detection accuracy were the main reasons for utilizing a skeleton detection model instead. Second, the relation between human pointing and objects in the scene needs precise camera calibration, such that the same object is referred to in both images. This can be circumvented by using a single camera for both visual perception tools, with RGB and depth perception functionalities.

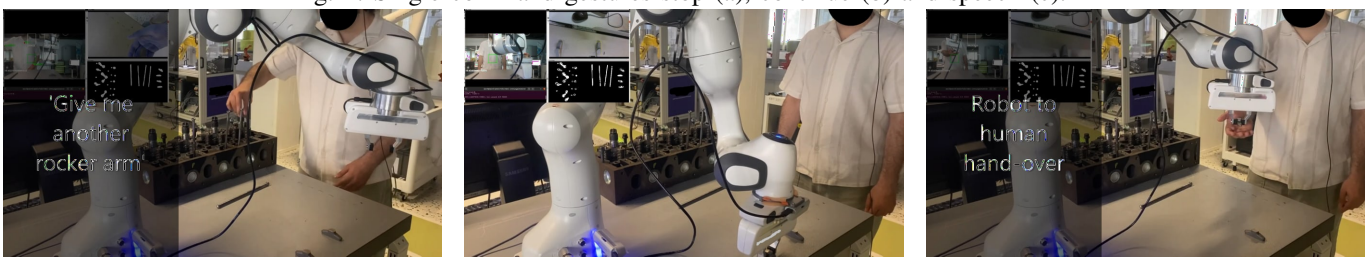
VI. Conclusions

This work investigated how multiple perception tools can be utilized and combined for effective human-robot collaboration. Human hand gestures and speech, as well as object detection, provide the input for robot actions, as coordinated by a person. Single modal perception serves to command basic robot actions (stop, continue) by gesture or speech. A co-speech gesture model is developed that combines human speech phrases, pointing gestures and object detection to command robot actions (pick and place, robot to human hand-overs) to specified objects in the scene. Experimental results demonstrate that co-speech gestures can be easily utilized for coordinating a shared task between human and robot.



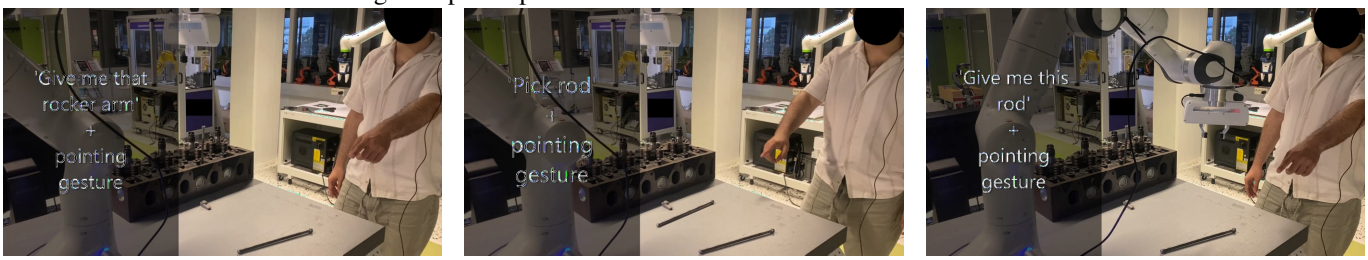
(a) (b) (c)

Fig. 4: Single command gestures stop (a), continue (b) and speech (c).



(a) (b) (c)

Fig. 5: Speech phrase to achieve robot to human hand-over.



(a) (b) (c)

Fig. 6: Co-speech gestures to achieve specified robot actions to objects.

REFERENCES

[1] N. Robinson *et al.*, "Robotic vision for human-robot interaction and collaboration: A survey and systematic review," *ACM Trans. Hum. - Robot Interact.*, vol. 12, no. 1, pp. 1–66, 2023.

[2] S. Gross and B. Krenn, "A communicative perspective on human–robot collaboration in industry: Mapping communicative modes on collaborative scenarios," *Int. J. of Social Robotics*, pp. 1–18, 2023.

[3] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, 11 2018.

[4] L. Wang *et al.*, "Symbiotic human-robot collaborative assembly," *CIRP Annals*, vol. 68, pp. 701–726, 2019.

[5] F. Semeraro, A. Griffiths, and A. Cangelosi, "Human–robot collaboration and machine learning: A systematic review of recent research," *Robot. Comput. Integr. Manuf.*, vol. 79, p. 102432, 2023.

[6] N. Sünderhauf *et al.*, "The limits and potentials of deep learning for robotics," *Int. J. Rob. Res.*, vol. 37, pp. 405–420, 4 2018.

[7] J. Fan, P. Zheng, and S. Li, "Vision-based holistic scene understanding towards proactive human–robot collaboration," *Robot. Comput. Integr. Manuf.*, vol. 75, p. 102304, 6 2022.

[8] A. Zacharaki, I. Kostavelis, A. Gasteratos, and I. Dokas, "Safety bounds in human robot interaction: A survey," *Safety Science*, vol. 127, p. 104667, 7 2020.

[9] T. Linder, N. Vaskevicius, R. Schirmer, and K. O. Arras, "Cross-modal analysis of human detection for robotics: An industrial case study," in *IEEE Int. Conf. Intell. Robots Syst.*, 9 2021, pp. 971–978.

[10] E. Magrini *et al.*, "Human-robot coexistence and interaction in open industrial cells," *Rob. Comp. Integr. Manuf.*, vol. 61, p. 101846, 2 2020.

[11] C. R. Qi *et al.*, "Frustum pointnets for 3D object detection from RGB-D data," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018.

[12] D. Osokin, "Real-time 2D multi-person pose estimation on CPU: Lightweight OpenPose," *arXiv preprint arXiv:1811.12004*, 2018.

[13] A. Angleraud *et al.*, "Coordinating shared tasks in human-robot collaboration by commands," *Front. Robot. AI.*, vol. 8, 10 2021.

[14] T. B. Ionescu and S. Schlund, "Programming cobots by voice: A human-centered, web-based approach," *Proc. CIRP*, vol. 97, pp. 123–129, 2021.

[15] A. Boteanu *et al.*, "A model for verifiable grounding and execution of complex natural language instructions," in *IEEE Int. Conf. Intell. Robots Syst.*, 2016, pp. 2649–2654.

[16] J. K. Behrens *et al.*, "Specifying dual-arm robot planning problems through natural language and demonstration," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2622–2629, 2019.

[17] P. Bremner and U. Leonards, "Efficiency of speech and iconic gesture integration for robotic and human communicators-a direct comparison," in *IEEE Int. Conf. Robot. Autom.*, 2015, pp. 1999–2006.

[18] H. Chen, M. C. Leu, and Z. Yin, "Real-time multi-modal human–robot collaboration using gestures and speech," *J. Manuf. Sci. Eng.*, vol. 144, no. 10, p. 101007, 2022.

[19] Alpha Cephei, "Vosk Speech Recognition Toolkit," https://github.com/alphacep/vosk-api, 2023.

[20] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[21] G. Sharma, R. Pieters, and A. Angleraud, "Engine assembly dataset," http://dx.doi.org/10.5281/zenodo.7669593, Feb. 2023.

[22] O. Mazhar *et al.*, "A real-time human-robot interaction framework with robust background invariant hand gesture detection," *Robot. Comput. Integr. Manuf.*, vol. 60, pp. 34–48, 12 2019.