Impact of verbal instructions and deictic gestures of a cobot on the performance of human coworkers

Rami Younes, Frédéric Elisei, Damien Pellier, Gérard Bailly

Impact of verbal instructions and deictic gestures of a cobot on the performance of human coworkers

Rami Younes
GIPSA-lab & LIG-Lab
Univ. Grenoble-Alpes
St Martin d'Hères, FR
0000-0001-9507-5598

Frédéric Elisei
GIPSA-lab
Univ. Grenoble-Alpes
St Martin d'Hères, FR
0000-0002-1295-3445

Damien Pellier
LIG-Lab
Univ. Grenoble-Alpes
St Martin d'Hères, FR
0000-0003-3791-8985

Gérard Bailly
GIPSA-lab
Univ. Grenoble-Alpes
St Martin d'Hères, FR
0000-0002-6053-0818

***Abstract*— This paper investigates the effectiveness and efficiency of incorporating pointing gestures as well as hand-speech synchronization policies into instruction delivery, as would be used in an industrial case with a cobot. Through brick assembly tasks, our study explores the integration of pointing gestures into human-robot interaction, extending prior research on verbal instruction efficacy. Results show that pointing gestures significantly reduce errors compared to verbal instructions alone, especially for complex tasks. However, this improvement comes at the cost of increased task completion time. We also show that depending on this synchronization, the user might delay its action until all information is presented instead of exploiting the information as it arrived. This study emphasizes the potential of pointing gestures and hand-speech synchronization in improving human-robot interaction and suggests further research for optimal integration.**

I. Introduction

Verbal expression of plans finds application across a wide spectrum of Human-Robot Interaction (HRI) scenarios. In our prior study [28], we investigated various verbal styles to determine their effectiveness in minimizing errors and reducing assembly time in a one-on-one instruction-based shared assembly task, where the robot acts as the primary instructor and the human as the learner.

Today's world features a wide range of robots, from industrial machines designed for specialized tasks to humanoid robots that can interact with humans more naturally and intuitively. Pointing gestures consist of closing all the fingers except the index finger which will serve as the pointer [8], [13]: such shape is easy to mimic by robots equipped with fingers. Our work is with an industrial robot (as shown in fig. 1), which uses grippers as the end effector. In this paper, we extend our previous work to study the impact of adding a non-verbal modality (pointing).

In addition to speech, gesturing is an efficient way of providing information during an interaction. Gestures are often categorized based on their role in communication (e.g. deictic, iconic, symbolic/metaphorical, beat gestures) [22]. With the use of both pointing and verbal modalities, we need to consider the coordination between hand movements and speech, which presents its own set of challenges. Indeed a robot producing gestures that do not match the rhythm or meaning of its speech can break down its interaction with the human. It was shown in [26] that including gestures along with speech in HRI rendered the robot more anthropomorphic and likable, and the participants preferred interacting with a robot using the two modalities instead of communicating through speech alone. Moreover [5] shed light on the challenge of the coordination between the pointing and the verbalization (speech-hand coordination). We will focus on the deictic gestures, to see how they can affect the communication of the plan to the human agent, as well as finding a good approach to the speech-hand coordination challenge.

This work is funded by MIAI (ANR-19-P3IA-0003).

Fig. 1. One-on-one human-robot interaction with YuMi ABB robot. In this assembly task, the robot mainly acts as the master instructor.

We start by questioning whether there is a need to introduce the pointing modality, especially when the content of the verbal instructions is unambiguous. Then, we aim to compare different hand-speech synchronization policies.

We have formulated several hypotheses that we aim to test in the context of a shared assembly task, where the primary goal is task completion, focusing on both effectiveness and efficiency. Our main concern lies in error reduction while maintaining a reasonable time frame for task completion.

Hypotheses:

- H1: pointing reduces the number of errors.
- H1-bis: pointing reduces the number of errors, for more difficult tasks.
- H2: pointing increases the time to complete.
- H2-bis: pointing helps reduce the time to complete more difficult tasks.
- H3: pointing improves human acceptability.
- H4: Human participants will exploit the information as soon as it is available.

To test these hypotheses, we designed an experiment following the context of our previous web-based experiment [28], featuring a shared LEGO™ assembly task where the robot serves as the primary instructor and the human as the learner (fig. 1). However, this experiment extends the previous one by employing an industrial robot to perform a sequence of continuous actions rather than independent standalone actions with a virtual robot depicted on screen.

This paper is organized as follows: Section 2 contains related work on verbal and non-verbal communication, namely pointing; Section 3 describes the overall architecture of our control model, with a closer look at pointing and verbalization; Section 4 introduces the setup of the experiment used to validate our hypotheses; finally, Section 5 presents the results of this multimodal interaction.

II. Related Work

In this section, we review several studies that explore the role of gestures and speech in improving HRI effectiveness.

Bangerter et al [6] conducted a study comparing different strategies for referring to objects within a shared visual space. They found that both pointing and verbal location description enhance accuracy, with pointing before speech leading to reduced consideration of items. Building on this notion of the importance of pointing, Holladay et al [17] proposed a mathematical framework for robots to generate pointing configurations optimizing the legibility of target objects. Their work emphasizes clarity over expense, enhancing interpretability, especially for novice users. Similarly, Haring et al [15] investigated the use of gaze and pointing gestures in humans following humanoid robot instructions, revealing the significance of pointing gestures in augmenting speech-based instructions. Expanding the scope to real-time interaction, Salem et al [27] developed a framework enabling robots to produce synthetic speech and gestures. Their study highlights the positive perception of robots displaying hand and arm gestures alongside speech. Alikhani et al [4] explored how robotic arms communicate task information through pointing actions, distinguishing between identifying objects and locations, thus contributing to the understanding of effective communication in collaborative robotics. Admoni et al [2] modulated speech, head, and pointing behaviors based on object identification ambiguity, demonstrating the effectiveness of non-verbal cues in high ambiguity scenarios. Further emphasizing the importance of non-verbal cues, Ali et al [3] evaluated the effectiveness of robot-provided directions, showing that non-verbal directions lead to quicker task completion compared to verbal directions. Considering human behavior, Liu et al [21] examined human pointing behaviors in various contexts, anticipating differences in precision based on the target and conversational openness. Finally, Gleeson et al [12] focused on explicit and intuitive gestural communication in industrial human-robot teams, highlighting the significance of context in gesture interpretation for accurate understanding.

When dealing with multiple modalities, a proper coordination is important. In terms of hand-speech coordination for deictic gestures, several studies have examined the influence of gesture production on the onset of associated verbal referent [10], [20]. They conclude that speech onset is delayed when a gesture is to be performed simultaneously with speech. In [18], they investigated how manual pointing gestures align with different aspects of speech, such as vowels, consonants, or tones. Initial findings indicate that pointing gestures tend to align more closely with tone gestures. In his thesis [24], Roustan studies the coordination between manual gestures and speech in the context of designation. Different types of gestures, including pointing, were compared in his work. He explores coordination in a more natural and interactive task and shows that when someone gestures with their hands, it often synchronizes with what they are referencing.

Our work builds on this work and focuses on advancing proximate human-robot interaction in shared assembly tasks, with an industrial robot assuming the role of an instructor. We emphasize investigating the effectiveness of incorporating robot pointing gestures alongside speech, compared to speech alone. Additionally, we explore different policies of hand-speech coordination to better understand their impact on interaction dynamics.

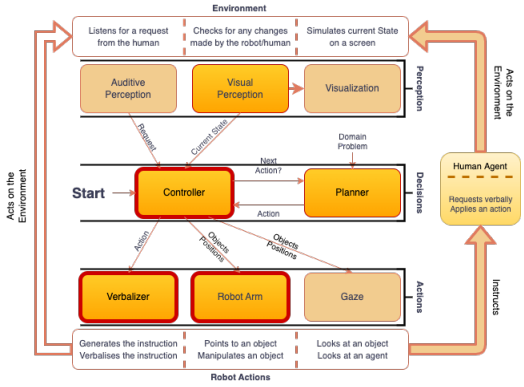

Fig. 2. Architecture of our HRI system. The components discussed in this paper are highlighted in red.

III. Architecture

This section follows the architecture from [28], focusing on speech (modifications added to the verbalizer module) and pointing modalities as well as their coordination (Fig. 2).

The current perception module is in charge of detecting the participant's pick and place gestures in the collaborative task. Note that the experimenter will manually confirm that the user has ended the execution of the instruction. And in case of an error made by the participant (i.e. incorrectly placed object), the experimenter here intervenes to fix the mistake in order to proceed with the experiment.

A. Verbalizer

A verbalizer is used for the generation of verbal instructions. We use the one described in [28] with the addition of certain tags. These tags help us identify certain parts (slices) in the instructions. In our tasks, we require three

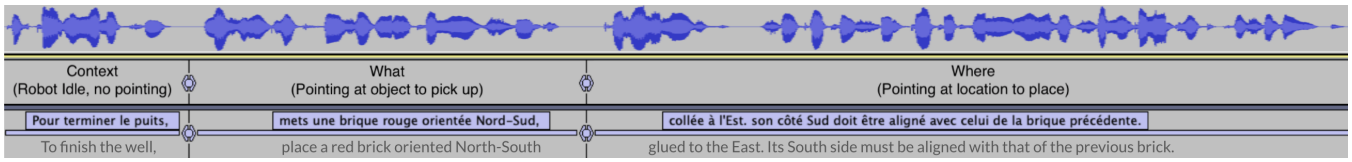

Fig. 3. Decomposition of an instruction

essential elements: context, objectives, and methodology to complete the task. Since we will recruit multiple modalities, such as the robot's arm movements, it's crucial to include information about them. In this case, in tandem with the robot's arm movements, we can easily insert tags for other use cases, gaze for instance. The tags we have used serve to categorize different aspects of the instructions, and they are as follows:

- ⟨Context⟩...⟨/Context⟩ As the name suggests, this tag provides the context, explaining the higher-level task we are working on.
- ⟨What⟩...⟨/What⟩ This tag references the object we are addressing.
- ⟨Where⟩...⟨/Where⟩ This tag contains information about the placement of the reference object that needs to be relocated.
- ⟨RETN⟩ This tag indicates a definitive retraction of the robot arm at the end of the instruction, returning the arm to its usual idle position.
- ⟨Rdv⟩ In some cases, we decided to tag keywords to facilitate hand-speech coordination (discussed in detail in section III-C).

Here is an example of a tagged instruction, in French: '⟨Context⟩Pour terminer le puits,⟨/Context⟩ ⟨What⟩mets une ⟨Rdv⟩brique rouge orientée Nord-Sud,⟨/What⟩ ⟨Where⟩collée à l'Est. Son côté Sud doit être aligné avec celui de la ⟨Rdv⟩brique précédente.⟨/Where⟩⟨RETN⟩'.

Figure 3 shows the English translation and an example of this slice decomposition, where we have 3 different slices of the audio instructions along with the corresponding robot gesture. The first slice is the *context* and the robot arm does not move – 'To finish the well,'. The second slice is *what* to pick and place – 'place a red brick oriented North-South'. Finally, the third is the location detailing *where* to place the object – 'glued to the East. Its South side must be aligned with that of the previous brick'. We can also see how the robot's movements are linked with each part of the instruction. In the second and third slices, the robot arm moves in order to point at the reference and target positions. Note that other slices and keywords could have been added to associate gestures and verbal content, in particular for cueing parts or properties of objects (e.g. 'côté Sud' *South side* with a flat hand).

Finally, for instruction synthesis, among multiple voices provided in [19], we chose the female voice of Nadine Eckert-Boulet (NEB). We use a synthesizer that is able to handle most - if not all - problematic mispronunciations in French (in particular handling homographs, liaisons, etc) as well as mixed text/phonetic input. It allows to add emphasis, pauses, and most importantly provides the duration of all phones of the utterances. The emphasis helps to focus the intonation or key parts of the instruction, in our case the object and/or target location. The pauses along with the computed time for each utterance allow us to slice an instruction according to synchronization requirements.

Combining our tagged-generated instructions with the features of our synthesizer is what allows us to coordinate verbal instructions with gestures.

B. Pointing gestures

Gesture generation

In the experimental procedure concerning robot gestures, we first position the robot arm at the designated location within the environment. Next, we utilize ABB's Robot Web Services to capture and record this precise position. Subsequently, using the MoveIt ROS node, we generate and store the trajectory for the robot's movement [7]. Afterward, adjustments are made to the trajectory file to ensure reasonable duration. Finally, the trajectory is executed through External Guided Motion (EGM), facilitated by a separate license [1]. The positions of the gestures were chosen to appear as human-like as possible to enhance the robot arm's ability to interact naturally with people. For the pick-and-place actions, a similar approach is adopted, involving positioning the robot arm and recording its location, followed by utilization of the RAPID interface for controlling the robot arms. The architecture allows for real-time gesture generation instead of storing them. Note that the time-consuming HTTP requests for acquiring the generated gestures pose challenges in maintaining a seamless, uninterrupted interaction.

Between instructions, we instruct the robot arms to retract (i.e. move away from the environment to an idle position). Thus giving a clear message that the instruction is over as well as giving the opportunity to the participant to perform the instruction.

Our gesture control policy includes two other features: (a) the robot points at the target objects/positions at a moderate distance ($\approx$ 14 cm); (b) when the robot arms retract, the grippers open up to signal an idle state. The grippers then close back as soon as the arm is ready to point. We believe that this will help the human agent anticipate that the robot is about to move without explicit verbal statements.

Balancing Speed and Precision in 3D Object Pointing

Among the key aspects that define the efficacy of robot movements, speed and accuracy stand out as paramount considerations, particularly when it comes to pointing and

interacting with three-dimensional objects. Each movement of the robot arm, whether it is reaching for an object, executing a task, or repositioning itself, is executed within a specific time frame [23]. This temporal aspect is critical as it directly influences the efficiency, safety, and overall effectiveness of the robot's actions. We take advantage of Fitts's law which predicts the time T required for a user to point to a target based on its distance D and target width W (in mm):

$$T = a + b.log_2(1 + \frac{D}{W}) \quad (1)$$

In their paper [23], the authors extend this law to three-dimensional pointing tasks. In our work, we set $a = 166$ and $b = 230$ (ms); b being the slope where with each unit increase in ID, the movement time goes up by about b ms, and a being the estimated minimum time needed for the simplest task $(ID = 0)$.

In scenarios where rapid and somewhat precise object manipulation is required, shorter movement durations might be preferred to enhance productivity. On the other hand, longer movement durations may be more appropriate for tasks demanding extreme precision, minimizing errors, and ensuring safety.

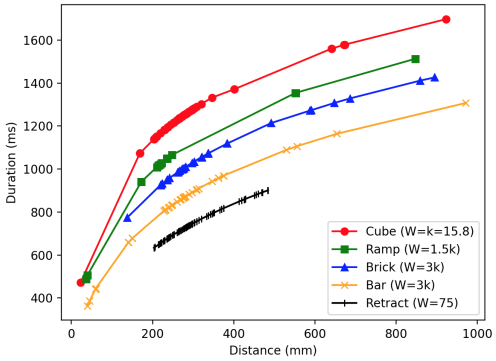

Fig. 4. Fitts duration (ms) wrt distance (mm) per target.

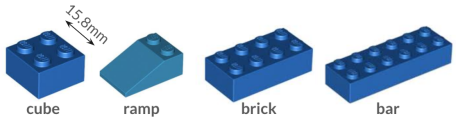

Fig. 5. LEGO™ dimensions

Decomposing all the actions of our scenarios, we identify five distinct classes per target (cube, ramp, brick, bar, retract): Fitts' Law is applied to these movements, revealing that greater distances result in shorter travel durations, while precise pointing for smaller objects leads to slower speeds and longer durations as presented in Fig. 4. The LEGO™ objects' dimensions are as shown in Fig. 5 ($W = 15.8$ mm for a LEGO™ cube), along with a larger target size ($W = 75$ mm) for the retraction. The width W of the target (LEGO™ object) corresponds to the length (in mm) of its longer side.

C. Hand-speech coordination

Having gestures with verbal instructions alone is insufficient; a seamless and comprehensible interaction demands a synchronization of both modalities. Figure 3 illustrated the auditory slices we use. With our synthesizer, we know their durations. We also master the corresponding robot movements and their respective durations. We will exploit these to build two different synchronizing policies:

- *ASAP* "As Soon As Possible" policy is a basic auto-play of both the speech and the corresponding gesture.
- *JIT* "Just In Time" policy involves synchronizing the two modalities to reach simultaneously a predefined target in the interaction. As humans, we synchronize the gesture apex with the onset of the accented syllable of the verbal referent [14], [25]. The JIT approach involves delaying the initiation of either the pointing trajectory or speech to ensure that both modalities align precisely at the intended moment.

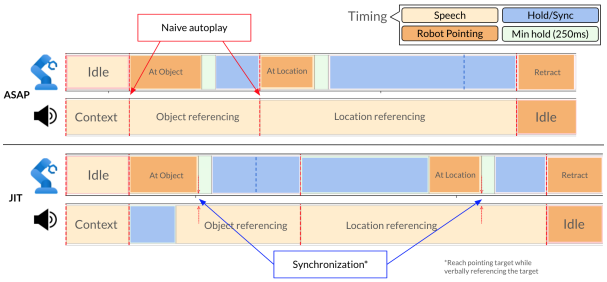

Fig. 6. Showing the difference between ASAP and JIT synchronization policies

Fig. 6 exemplifies the difference between the two policies. Following the example in Fig. 3, which comprises three verbal slices (context, object target, and location target), where context involves no pointing gesture. Consequently, in both ASAP and JIT conditions, the robot utters contextual information without an accompanying gesture. The ASAP policy dictates that both pointing gestures and verbal slices start simultaneously, disregarding the temporal target information in the speech.

Conversely, under the JIT policy, we delay either the speech (by adding silence) or gesture for the second slice, ensuring that the pointing of the targeted object synchronizes with the onset of the verbal reference. Similarly, for the third slice, the referencing of the relative positioning object is synced with the location pointing target. For a comparison of the two policies, please play the accompanying video[1]. Ultimately, in both policies, the decision is made to allow the modality that completes first to wait for the other, constraining a minimum hold time of 250 ms for pointing gestures.

The next section outlines the experimental protocol used in this study. We detail the specific steps taken to conduct the experiment.

[1]Video: `https://youtu.be/QoNLrh6qixk`

IV. Experiment

In order to study how pointing and hand-speech coordination affect human-robot collaboration, we create an experiment with an ABB industrial robot and a LEGO™ table as the working environment (see fig. 1). The robot is ambidextrous The robot is ambidextrous, and each arm is chosen based on the shortest distance to the target.

Participants are asked to observe the robot as it positioned the first LEGO™ piece on the game board. They will place all the other pieces, as instructed by the robot. A total of 11 scenes are utilized, comprising 2 example scenes and 9 test scenes with randomized order and conditions. Fig. 7 shows the scenes used, featuring diverse colors, types, structures, and positions, generating ambiguity and challenges to evaluate human-robot collaboration. The number of LEGO™ pieces to be placed falls between 4 and 13.

Descriptions of verbal instructions and scenes can be found here.

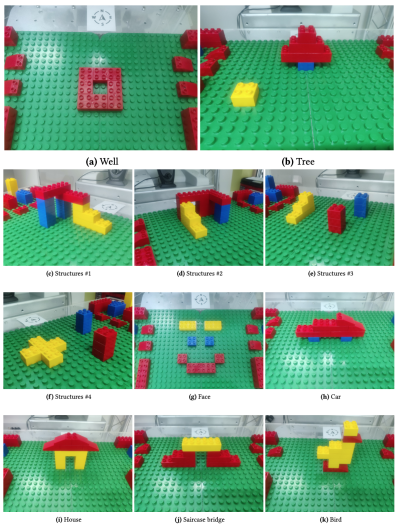

Fig. 7. Scenes (objects) to be assembled (a-b for familiarisation)

30 French speakers were recruited from the University of Grenoble Alpes. 90% are right-handed, and 23.33% identify as men. None of whom had previously participated in an HRI experiment. They were required to accurately place the correct element as instructed, with the least amount of mistakes, and as fast as possible. (see video[2]).

We compared 3 instruction methods :

- **AUDIO**: no pointing i.e. verbal instructions only
- **ASAP**: As Soon As Possible speech/hand triggering
- **JIT** Just In Time coordination

In the study described, each participant went through a familiarization phase consisting of two scenes with gesturing. This phase was randomized per participant to get a balanced distribution between the order and the choice of instruction method (ASAP, JIT) for the two scenes. After the familiarization phase, the 9 testing scenes are randomly divided into three blocks, with each block corresponding to one instruction method: AUDIO, ASAP, and JIT. The order of the familiarization scenes and testing scenes was evenly distributed among the participants. This distribution allowed each scene to be tested 10 times per instruction method, resulting in a total of 30 tests overall.

[2]Video: `https://youtu.be/hduHFnn_njw`

A. Measurements

The subjective evaluation is done with 5 questionnaires, each on a 5-point Likert scale (modified version from [11]). The same questionnaire is given before the beginning and at the end of the experiment to collect participant profile data and to assess how the experiment influences their initial responses. This assessment is made by comparing the responses to identical starting and concluding questions. Each of the remaining three questionnaires is filled at the end of each condition block, to gather data on each instruction method. This includes the NASA Task Load Index (NASA-TLX) [16], which is the most common, subjective, multi-dimensional framework [9] to measure the cognitive load. The objective is to assess and compare the difficulty and efficiency of the three models.

For the objective evaluation, we measure (a) the number of errors during the interaction (i.e. choosing the wrong object, pick-up place or placement); (b) different durations between multimodal time marks. Fig. 8 shows the different time-marks for a standard instruction and table I defines the durations based on these marks. For our measurements, we consider that the human pick-up occurs when they make contact with the last object they pick up, while the put-down happens when they confidently place the object and remove their contact from it. Robot retraction occurs the moment its arm is retracted from the workspace, and robot pointing is when the end effector reaches the target. Finally, object mention refers to the moment the object to be picked up is verbally referred to.

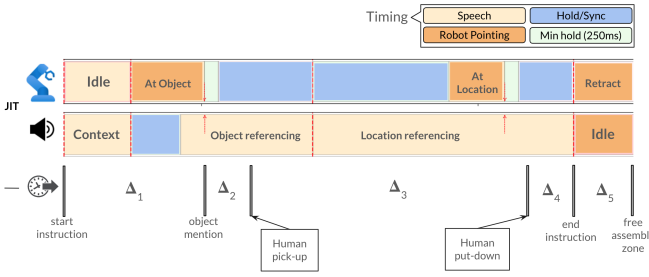

Fig. 8. time-marks for a standard instruction

It is important to note that the positions of the human pick-up and put-down time marks as presented in figure 8 are illustrative. Looking at the instruction in figure 3, it is possible that the human anticipates the object and/or location, or even decides to wait until the end of the instruction before manipulating the environment.

V. Results

We test our hypotheses from section I for two main comparisons; (1) AUDIO vs. {ASAP, JIT} (i.e. with/without

TABLE I

DURATIONS BASED ON TIME-MARKS FROM FIG. 8

Duration	Value
Pick-up	$\Delta_{1,2}$
Time to complete	$\Delta_{1,3}$
Instruction	$\Delta_{1,4}$
Object anticipation	Δ_2
Pick and place	Δ_3
Pick up after robot retraction	$\Delta_{3,4}$
Location anticipation	$\Delta_{4,5}$

$^*\Delta_{i,j} = \sum_{k=i}^{j} \Delta_k$

pointing) and (2) ASAP vs. JIT (i.e. comparing synchronization policies).

We categorize instructions by difficulty. It varies based on factors like object attributes and task requirements. We first define "stud_shift", representing the cumulative count of studs needed for spatial manipulation, such as "4 down and 2 to the right" for a stud_shift of 6. An *easy* instruction involves no changes in the object's type or color and requires no stud_shift, meaning no counting of studs is needed. Conversely, a *complex* instruction includes additional information or a stud_shift greater than 2, indicating the need to count multiple studs in different directions. By additional information, we mean the use of complementary details required by a more complex instruction. *Normal* difficulty falls between these extremes.

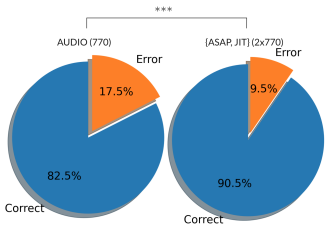

Fig. 9. Error %: AUDIO vs. {ASAP, JIT} (p-value < 0.001: ***)

A. Error analysis

Fig. 9 confirms **H1**: ASAP and JIT reduce the number of errors compared to AUDIO. Both statistical tests conducted, namely the Student's t-test and the Mann-Whitney U test, yielded extremely small p-values ($< 10^{-7}$).

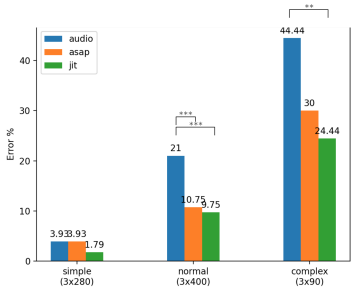

Fig. 10. Error %: AUDIO vs. ASAP vs. JIT per task difficulty

Fig. 10 confirms **H1-bis**: including pointing reduces the number of errors for complex tasks, but only JIT compared to Audio. ASAP and JIT reduce the number of errors compared to Audio. However, only the statistical tests (the Student's t-test and the Mann-Whitney U test) conducted on AUDIO vs. JIT yielded small p-values (0.002). This also confirms that JIT, compared to ASAP, reduces the number of errors for more complex tasks. Thus we establish that the Audio instruction method produces the highest number of errors overall compared to ASAP and JIT and that JIT reduces a more significant number of errors compared to ASAP, especially for more complex tasks.

TABLE II

ERROR RATIO ON TABLE VS. ON STRUCTURE ACTIONS PER CONDITION

	Audio	ASAP	JIT
on table (3 x 190)	28.95%	17.37%	13.16%
on structure (3 x 580)	13.79%	8.28%	7.07%

Finally, we point out that error ratios decrease progressively from Audio to ASAP to JIT for 'on table' and 'on structure' action types (tab. II). Specifically, actions 'on table' have higher error ratios compared to actions 'on structure'. This indicates that JIT is the most effective, yielding the lowest error ratios, followed by ASAP and then Audio. Additionally, performing 'on structure' actions significantly reduces error ratios compared to 'on table', suggesting a notable improvement in accuracy when actions are structured.

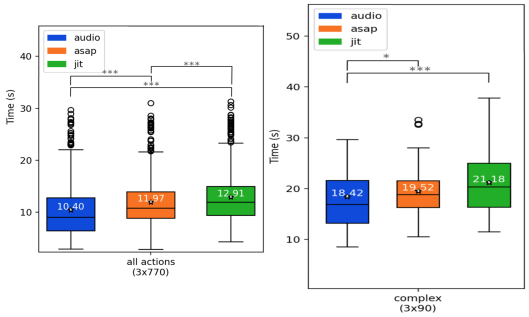

Fig. 11. Time to complete: AUDIO vs. ASAP vs. JIT (Left: for all actions — Right: for complex actions)

B. Timing analysis

Fig. 11 confirms **H2**: ASAP and JIT increase the time to complete compared to AUDIO. And JIT increases the time to complete compared to ASAP. Both the Student's t-test and the Mann-Whitney U test yielded small p-values, ranging from 10^{-5} to 10^{-24}

Fig. 11 refutes **H2-bis**: ASAP and JIT increase the time to complete compared to AUDIO for complex tasks and similarly for JIT compared to ASAP. Both the Student's t-test and the Mann-Whitney U test yielded small p-values. ($\approx 10^{-2}$ and $\approx 10^{-4}$).

We see that regardless of the task difficulty, including pointing will increase the time to complete compared to standalone verbal instructions.

Finally, we compare the durations of object placement after clearing the assembly zone, i.e. Δ_5 in Table I. Both

ASAP and JIT were affected by the robot arm being held in the assembly zone. Participants decided to pick up the object as soon as the arm starts to retract, JIT being faster than ASAP. The data from the experiment reveals that participants did not anticipate object nor location for most actions. This disconfirms **H4**: human participants waited for all the information before exploiting it. This could also be related to why we are not able to confirm **H2-bis**.

TABLE III

SIGNIFICANT DIFFERENCES IN QUESTIONNAIRE

Affirmation	Method1	Method2	p-value
A1. Vous avez été gêné(e) lorsque le robot a pointé les objets *You were bothered when the robot pointed to the objects*	JIT (1.60)	ASAP (1.80)	<.0001
A2. Le pointage était bien synchronisé avec la parole *The pointing was well synchronized with the speech*	ASAP (4.60)	JIT (4.67)	<.0001
A3. Le pointage des objets par le robot était précis *The robot's pointing at the objects was precise*	ASAP (3.87)	JIT (4.20)	.0444
A4. Vous avez réussi à accomplir ce qu'on vous a demandé de faire *You managed to accomplish what you were asked to do*	Audio (3.17)	JIT (3.73)	.0296
A5. La tâche vous a demandé beaucoup de concentration *The task required a lot of concentration from you*	ASAP (3.73) JIT (3.77)	Audio (4.13)	.0363 .0380
A6. Le rythme des instructions était trop rapide *The pace of the instructions was too fast*	JIT (2.53)	Audio (3.20)	.0042
A7. L'interaction avec le robot vous a paru fluide et prévisible *The interaction with the robot seemed smooth and predictable to you*	Audio (3.40)	JIT (3.97)	.0265
A8. Vous étiez stressé *You were stressed*	JIT (2.23)	Audio (2.67)	.0376
A9. La tâche vous a demandé beaucoup d'effort *The task required a lot of effort from you*	ASAP (3.13) JIT (2.90)	Audio (3.53)	.0462 .0016

C. Qualitative evaluation

In this analysis, we employed ordinal logistic regression to examine the relationship between various response variables and experimental conditions. For each response variable, we fitted a cumulative link mixed model (CLMM) using the `clmm` function, which accommodates the ordinal nature of the data and includes a random intercept for the `participant id` variable to account for subject-specific variability. Multiple comparisons between conditions were performed using the `emmeans` package to obtain pairwise contrasts. We only present the results for the affirmations holding significant pairwise comparisons in table III. Participants experienced significantly less discomfort and stress, better synchronization and precision in the robot's actions, and an overall smoother and more predictable interaction with JIT. JIT also required less concentration and effort from participants, indicating higher efficiency and user-friendliness. Hands-on experience with the collaborative robot (cobot) during the experiment notably increased participants' confidence and trust in the cobot's information and reliability. These findings highlight the benefits of JIT in improving user comfort, efficiency, and trust in human-robot interactions.

With these affirmations, not only do we confirm **H3** that including pointing improves human acceptability, but also that the JIT method consistently outperforms both the ASAP and Audio policies in different aspects.

VI. DISCUSSION

In this section, we address the limitations of our experiment and suggest areas for improvement.

Maximum hold duration

As previously discussed, participants waited for the robot to retract its arm before manipulating the environment (pick and place). While a minimum hold was introduced to ensure a minimum 250 ms pointing hold even when speech was too short, no maximum hold was set. We believe that the absence of a maximum hold for the robot arm may have contributed to the observed lack of object/location anticipation. Currently, the implemented hold keeps the robot arm engaged until the speech ends. This raises several issues, particularly for longer audio inputs (such as sentences). Firstly, the prolonged presence of the robot arm can obstruct the working environment. Secondly, it disrupts the flow of interaction, as it may force the human to wait unnecessarily, especially when actions can be anticipated.

Error correction guidance

When errors occur during human-agent interaction, made by the human participant, it's essential to consider the implications and limitations of the robot's ability to address them. While the robot may possess some capacity to rectify errors, it's not feasible for now to cope with all situations. Relying solely on the robot to physically correct errors introduces several challenges.

Firstly, involving external human intervention, often in the role of a corrector to rectify errors, may disrupt the flow of interaction for the human participant. This interruption can lead to disengagement from the working environment and may adversely affect the participant's mind (such as arising from social judgment). Furthermore, the reliance on external intervention to rectify errors can introduce inefficiencies and delays in the task execution process.

To mitigate these challenges and ensure smoother interaction dynamics, it may be beneficial to include a correction policy to guide the human agent – specific to AUDIO, ASAP, or JIT conditions – on how they can correct the mistake.

Updating verbal instructions for use with pointing

We use identical verbal instructions across all three conditions (Audio, ASAP, JIT). This decision was motivated by the desire to avoid the risk of changing behavior due to the verbal content and to minimize differences in duration. For instance, in the Audio condition, an example instruction would be "Take a red brick oriented East-West"

However, we made minor adjustments during the generation step for the pointing conditions (ASAP and JIT), such as using "this" instead of "a" to better align with the pointing gesture. Consequently, the instruction used in the ASAP/JIT conditions was, "Take this red brick oriented East-West"

It is possible to further streamline the speech time, particularly in the pointing conditions, without compromising clarity or effectiveness. Therefore, to minimize the overall time added by the pointing methods and aim to match the duration of the faster AUDIO condition, we propose to reduce redundancy wherein the verbal instruction for the ASAP/JIT conditions; e.g. simplifying to "Take this brick" omitting unnecessary details that can be inferred through the pointing gesture. This adjustment not only reduces the duration of speech but also enhances the efficiency and fluidity of the interaction process when coupled with the pointing gestures.

VII. Conclusion and perspectives

In this paper, we explore the integration of non-verbal pointing gestures in an HRI scenario, building upon our previous work on verbal instruction styles' efficacy. We hypothesize that including pointing gestures will lead to error reduction, albeit potentially increasing task completion time. Furthermore, we anticipate that participants will exploit available information promptly and that the inclusion of pointing will enhance overall human acceptability. The inclusion of pointing gestures significantly reduces errors compared to verbal instructions alone, particularly for more challenging tasks. However, this reduction in errors is accompanied by an increase in task completion time, regardless of task complexity. The qualitative analysis found that the JIT method consistently outperforms both ASAP and Audio methods across various dimensions. To our surprise, participants did not exploit available information promptly, waiting until all information was presented, and the robot arm was retracted before taking action. Future work may focus on integrating gaze into our HRI, which could significantly enhance interaction dynamics. By incorporating gaze information, we introduce the need to coordinate the gaze with the current modalities However, this addition will allow robots to establish joint attention with human participants, leading to more effective communication and collaboration. Additionally, integrating obstacle avoidance capabilities into motion planning algorithms holds promise for improving HRI performance. However, achieving a balance between obstacle avoidance and maintaining natural arm movements would still need to be addressed. Finally, interactions could extend involving multiple users. This entails addressing the complexities associated with planning and coordinating interactions between the robot and multiple users, including managing competing instructions, individual preferences, and collaborative decision-making processes.

References

[1] ABB Externally Guided Motion EGM. `http://new.abb.com/products/ABB.PARTS.SEROP3HAC054376-001`, 2017.

[2] Henny Admoni, Thomas Weng, and Brian Scassellati. Modeling communicative behaviors for object references in human-robot interaction. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3352–3359. IEEE, 2016.

[3] Waqar Ali and Andrew B. Williams. Evaluating the effectiveness of nonverbal communication in human-robot interaction. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 99–100, 2020.

[4] Malihe Alikhani, Baber Khalid, Rahul Shome, Chaitanya Mitash, Kostas Bekris, and Matthew Stone. That and there: judging the intent of pointing actions with robotic arms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10343–10351, 2020.

[5] Gérard Bailly and Frédéric Elisei. Speech in action: designing challenges that require incremental processing of self and others' speech and performative gestures. In *Natural Language Generation for Human-Robot Interaction (NLG4HRI)*, 2020.

[6] Adrian Bangerter and Max M. Louwerse. Focusing attention with deictic gestures and linguistic expressions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 27, 2005.

[7] Sachin Chitta. Moveit!: an introduction. *Robot Operating System (ROS) The Complete Reference (Volume 1)*, pages 3–27, 2016.

[8] Roberto Cipolla and Nicholas J. Hollinghurst. Human-robot interface by pointing with uncalibrated stereo vision. *Image and vision computing*, 14(3):171–178, 1996.

[9] L. Colligan, H. Potts, Chelsea T. Finn, and R. A. Sinkin. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. *International journal of medical informatics*, 84 7:469–76, 2015.

[10] Pierre Feyereisen. The competition between gesture and speech production in dual-task paradigms. *Journal of memory and language*, 36(1):13–33, 1997.

[11] Étienne Fournier, Christine Jeoffrion, Belal Hmedan, Damien Pellier, Humbert Fiorino, and Aurélie Landry. Human-cobot collaboration's impact on success, time completion, errors, workload, gestures and acceptability during an assembly task. *Applied Ergonomics*, 119:104306, 2024.

[12] Brian Gleeson, Karon MacLean, Amir Haddadi, Elizabeth Croft, and Javier Alcazar. Gestures for industry intuitive human-robot communication from human observation. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 349–356. IEEE, 2013.

[13] Mehmet Gokturk and John L. Sibert. An analysis of the index finger as a pointing device. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems*, pages 286–287, 1999.

[14] Chloe Gonseth, Anne Vilain, and Coriandre Vilain. An experimental study of speech/gesture interactions and distance encoding. *Speech communication*, 55(4):553–571, 2013.

[15] Markus Häring, Jessica Eichberg, and Elisabeth André. Studies on grounding with gaze and pointing gestures in human-robot-interaction. In *International Conference on Social Robotics*, pages 378–387. Springer, 2012.

[16] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[17] Rachel M. Holladay, Anca D. Dragan, and Siddhartha S. Srinivasa. Legible robot pointing. In *The 23rd IEEE International Symposium on robot and human interactive communication*, pages 217–223. IEEE, 2014.

[18] Jelena Krivokapic, Mark K. Tiede, Martha E. Tyrone, and Dolly Goldenberg. Speech and manual gesture coordination in a pointing task. In *Proceedings of Speech Prosody*, pages 2016–255, 2016.

[19] Martin Lenglet, Olivier Perrotin, and Gérard Bailly. Speaking rate control of end-to-end tts models by direct manipulation of the encoder's output embeddings. In *Interspeech 2022*, pages 11–15. ISCA, 2022.

[20] Willem J.M. Levelt, Graham Richardson, and Wido La Heij. Pointing and voicing in deictic expressions. *Journal of memory and language*, 24(2):133–164, 1985.

[21] Phoebe Liu, Dylan F. Glas, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. A model for generating socially-appropriate deictic behaviors towards people. *International Journal of Social Robotics*, 9(1):33–49, 2017.

[22] David McNeill. Hand and mind1. *Advances in Visual Semiotics*, 351, 1992.

[23] Atsuo Murata and Hirokazu Iwase. Extending fitts' law to a three-dimensional pointing task. *Human movement science*, 20(6):791–805, 2001.

[24] Benjamin Roustan. *Étude de la coordination gestes manuels/parole dans le cadre de la désignation*. PhD thesis, Université de Grenoble, 2012.

[25] Benjamin Roustan and Marion Dohen. Gesture and speech coordination: The influence of the relationship between manual gesture and speech. In *Interspeech*, pages 498–501, 2010.

[26] Maha Salem, Friederike Eyssel, Katharina J. Rohlfing, Stefan Kopp, and Frank Joublin. To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *Int. J. Soc. Robotics*, 5(3):313–323, 2013.

[27] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4(2):201–217, 2012.

[28] Rami Younes, Gérard Bailly, Frédéric Elisei, and Damien Pellier. Automatic verbal depiction of a brick assembly for a robot instructing humans. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 159–171, 2022.