Demonstrating Language-Grounded Control for Coordinated Robot Motion and Speech #277

Ravi Tejwani
Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02142 USA
Email: tejwanir@mit.edu

Chengyuan Ma
Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02142 USA
Email: macy404@mit.edu

Paco Gomez-Paz
Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02142 USA
Email: pjgomez@mit.edu

Paolo Bonato
Physical Medicine and Rehabilitation
Harvard Medical School
Charlestown, MA 02129 USA
Email: pbonato@mgh.harvard.edu

Harry Asada
Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02142 USA
Email: asada@mit.edu

***Abstract*—Recent advancements have enabled human-robot collaboration through physical assistance and verbal guidance. However, limitations persist in coordinating robots' physical motions and speech in response to real-time changes in human behavior during collaborative contact tasks. We first derive principles from analyzing physical therapists' movements and speech during patient exercises. These principles are translated into control objectives to: 1) guide users through trajectories, 2) control motion and speech pace to align completion times with varying user cooperation, and 3) dynamically paraphrase speech along the trajectory. We then propose a Language Controller that synchronizes motion and speech, modulating both based on user cooperation. Experiments with 12 users show the Language Controller successfully aligns motion and speech compared to baselines. This provides a framework for fluent human-robot collaboration. The implementation of the Language Controller framework is openly available as open source code to enable extension, verification, and reuse of the methods[1].**

I. INTENDED DEMONSTRATION

The demonstration will involve having audience volunteers sit beside a UR robot [1] equipped with the Language Controller. The robot will physically guide the volunteer's hand along a predefined trajectory on the table, representing a therapeutic exercise motion. Concurrently, the robot will provide verbal instructions and guidance to the volunteer through a speaker, using pre-recorded audio clips. As the robot moves the volunteer's hand, the volunteer will be encouraged to arbitrarily vary the amount of resistance they apply against the motion. The Language Controller will react in real-time to modulate both the pace of physical motion and the speech rate to maximize alignment. Audience members will clearly observe the robot slowing and speeding up the motion and audio in correlation with changes in applied resistance. Additionally, the phrase content and length of the speech instructions will adaptively change as the pace varies, demonstrating the controller's paraphrasing capabilities. Audience members will see that slower motions induce the use of longer instruction phrases, while faster motions result in shorter phrases.

[1]Github repository: https://github.com/language-playback-robot-controller/language-playback-robot-controller

II. INTRODUCTION

Robots have been enabled to collaborate with humans by providing physical assistance as well as verbal guidance during collaborative tasks. Research on robots providing physical assistance has shown robots assisting with heavy lifting and materials handling in warehouses and factories [2]; handling payloads, reducing physical strain on human workers [3]; and helping turn and lift patients, freeing up human nurses for other critical care tasks [4]. In addition to providing physical assistance, robots have also been used to provide verbal instructions and dialog interaction in human-robot collaboration [5], [6]. Using natural language capabilities, robots understand commands, ask clarifying questions, and provide guidance to human partners [7]. Research suggests that language-enabled robots lead to higher perceived collaboration quality compared to silent robot partners [8].

Human-robot collaboration remains limited by the lack of natural coordination between physical interactions and verbal communication. Simply combining robot's motions and its speech in parallel cannot achieve natural, fluent coordination. An open challenge remains to develop adaptive control frameworks that closely coordinate a robot's physical motions and speech utterances, dynamically modulating both based on real-time changes in human behavior.

We propose a Language Controller that synchronizes the robot's motion and speech under changing user cooperation by aligning the duration of the speech to that of the motion. Like human demonstrations, longer motion induces slower, longer speech and vice versa. The controller paraphrases speech utterances along the trajectory to match changing cooperation and speed. It does so by varying admittance parameters, audio pace, and adaptive paraphrasing. The controller is inspired by

principles derived from analysis of human-human physical interactions, specifically from an observational study of a physical therapist collaborating with a patient during therapeutic exercises at a rehabilitation center. From observations of the therapist's physical and verbal guidance, we identify core principles and translate them into formal control objectives for the controller (details in section IV).

We make the following contributions:

1) Formalize core principles to derive control objectives for natural human-robot collaboration – adaptive pacing, aligned speech-motion timing, correlating speech complexity with motion speed – from observations of human-human physical therapy interactions;
2) Language Robot Controller from the derived control objectives in order to align the robot motion with the verbal speech in the human-robot interaction;
3) Extensive human experiments that validate the controller and demonstrate its ability to align the pace of the robot motion with its speech.

III. RELATED WORK

A. Language Grounding for Robot Instructions

Prior work has focused on grounding natural language instructions to enable robots to follow commands, including techniques for mapping instructions to internal representations and actions [9]–[11]. Other efforts have targeted collaborative grounding of language between humans and robots for situated dialog and interactions [12], [13]. While enabling planning and collaboration, integrating robot physical motions and speech grounded in real-time human responses still remains an open problem. Our work aims to address this gap.

B. Variable Impedance and Admittance Control

Research has explored variable impedance and admittance control for safer and adaptive human-robot interaction, including dynamic modulation based on cooperation [14], adaptive admittance using EEG feedback [15], and online impedance variation for performance/safety trade-offs [16]. We incorporate admittance methods for compliant motion, but extend standard admittance frameworks by explicitly coupling the modulation of control parameters to the speech state. This ties motion control to verbal communication.

C. Language and Motion Integration

Recent works have combined language understanding with robotic planning and control, including mapping commands to executable specifications [17], [18] and leveraging implicit information to improve plan execution [19], [20]. However, physical motions of the robot and concurrent speech have not been integrated based on mutual understanding and real-time bidirectional communication. Our Language Controller addresses this by developing a control framework that coordinates motions and utterances grounded in human responses.

Fig. 1: Physical therapist is seen performing 'shoulder external rotation' therapy on to the patient with varying levels of physical resistance. The physical motions and speech data was recorded across different sessions.[2]

IV. PRINCIPLES AND CONTROL OBJECTIVES

Through the observational study of therapist-patient exercises at Spaulding Rehabilitation Hospital (Fig. 1), we identified the following core principles:

- The therapist planned the trajectories for each session, demonstrating the path before starting an exercise;
- The therapist adapted the pace of motions based on patient responses. When the patient struggled, she slowed down and gently guided them along the trajectory;
- The therapist aligned her speech with physical actions. She began verbal guidance at the start of motions and finished speaking around the end;
- Her speech rate and sentence length correlated with her physical motion speed - slower motions had slower, longer speech; faster motions had faster, shorter speech.

We derived the following formal control objectives from these principles. The robot must:

1) Guide the user through a predefined trajectory while modulating its velocity in response to user cooperation. High cooperation (low resistance) must lead to faster motion and Low cooperation (high resistance) must lead to slower motion;
2) Control the pace of its speech to maximize the alignment with its motion while adapting to varying user cooperation. We formalize alignment as a monotonic relationship between speech and motion duration (see V-C).
3) Paraphrase (choice of words) its speech dynamically along the trajectory, adapting to changing user cooperation. It must use shorter sentences under faster motion and longer sentences under slower motion.

Building upon the outlined control objectives, we formalize the Language Controller.

[2]Detailed therapy sessions videos can be seen at https://language-playback-robot-controller.github.io/therapy-sessions/

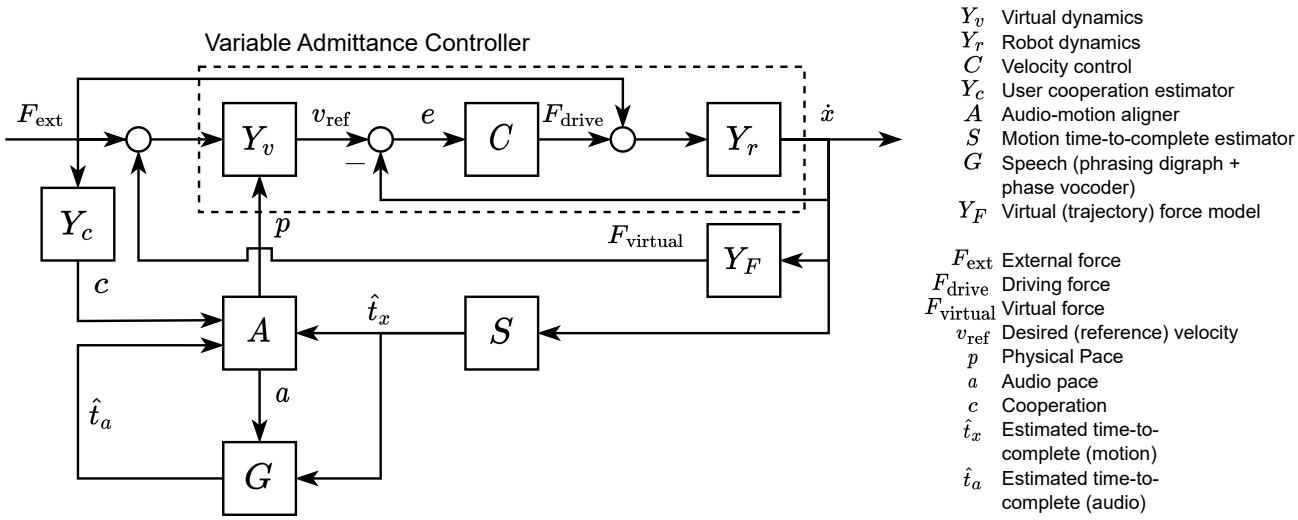

Fig. 2: Control scheme of Language Controller. External force F_{ext} (from user) and virtual force F_{virtual} (based on position to guide user through a predefined trajectory) are passed to the virtual dynamics Y_v to generate reference velocity v_{ref} for velocity controller C. C outputs a force that, together with F_{ext}, acts on robot dynamics Y_r. The resulting motion is given by $\dot{x}$. Motion time-to-completion estimator S uses $\dot{x}$ to estimate the time-to-completion (ETC) of the trajectory $\hat{t}_x$. Speech module G reports an ETC of the speech, $\hat{t}_a$, and cooperation model Y_c computes cooperation c from F_{ext}. $\hat{t}_x, \hat{t}_a$ and c are fed to audio-motion aligner A to update Physical Pace p and Audio Pace a. Physical Pace p changes the pace of the robot's motion by varying Y_v's parameters. Audio Pace a changes the pace of audio through a phase vocoder in G.

V. LANGUAGE CONTROLLER

A. Overview

Language Controller (Fig. 2) employs variable admittance control for robot motion and modulates the pace of motion and speech to maximize the alignment. The controller is designed to align the motion of the robot with its speech. It does so by updating "Physical Pace" and "Audio Pace" from estimated time-to-completions for both the robot's trajectory and its speech. The controller dynamically updates the paces based on real-time user response. Furthermore, the controller incorporates adaptive paraphrasing to modulate the speech content. It traverses a phrase graph representation to select appropriate wording and phrase length that matches the ideal duration of speech corresponding to the expected duration of motion.

B. Admittance Model

Admittance and Impedance Control are the two primary control schemes used in human-robot interactions [21], [22]. Humans actively control their limbs and resist unexpected movements, which positions humans as an impedance, necessitating robots to be treated as an admittance. We thus use Admittance Control, which converts external force F_{ext} into desired velocity v_{ref} via a virtual dynamics model:

$$M_0\dot{v}_{\text{ref}} + D_0 v_{\text{ref}} = F_{\text{ext}}. \tag{1}$$

where v_{ref} is the desired velocity. To lead the user through a predefined trajectory T, we extend the virtual dynamics model above with a virtual force F_{virtual},

$$M_0\dot{v}_{\text{ref}} + D_0 v_{\text{ref}} = F_{\text{ext}} + F_{\text{virtual}}. \tag{2}$$

F_{virtual} is dependent on the end effector position x and its closest point on the trajectory (x_d), defined as follows,

$$x_d = T(d), \text{ where } d = \arg\min_{0\le d\le 1}\|T(d) - x\|, \tag{3}$$

where we see the trajectory T as a directed curve $[0,1] \mapsto S$ where S is the state space. F_{virtual} consists of two parts:

$$F_{\text{virtual}} = \underbrace{K(x_d - x)}_{F_{\text{guide}}} + \underbrace{\|F_{\text{propell}}\|\mathbf{b}}_{F_{\text{propell}}}, \tag{4}$$

$$\text{where } \mathbf{b} = \lim_{d'\to d^+} \frac{T(d') - x_d}{\|T(d') - x_d\|}, \tag{5}$$

where F_{guide} leads the user back on track if they deviate and F_{propell} leads the user to complete the trajectory. This achieves our first objective that our robot should lead the user through the trajectory. In frequency domain, (2) can be expressed as

$$v_{\text{ref}} = A \cdot (F_{\text{ext}} + F_{\text{virtual}}) \text{ where } A = \frac{1}{M_0 s + D_0}. \tag{6}$$

We elaborate in V-E how to vary A with Physical Pace p to tune the behavior of this admittance model.

The virtual admittance model produces v_{ref} that is fed to the velocity controller C, which produces a driving force F_{drive} with actuators. The equation of motion of our robot is:

$$M_{\text{robot}}\ddot{x} = F_{\text{ext}} + F_{\text{drive}} = F_{\text{ext}} + C(v_{\text{ref}} - \dot{x}). \tag{7}$$

C. Defining Audio-motion Alignment

As was stated in IV, we observed from the demonstrative therapy sessions that the therapist began verbal guidance at the start of the motions and finished speaking around the end. Consequently, longer and slower motion is accompanied by

longer and slower speech and vice versa. Formally, the ideal duration of the speech, t_a, is an increasing function of the duration of the motion, t_x,

$$t_a = f(t_x), \text{ where } f'(\cdot) > 0. \tag{8}$$

We refer to the model f that correlates the two durations as **the alignment model**. f can vary between different contexts. In this demonstration, since we observed from the therapy videos that the speech guidance is approximately as long as the motion itself, we set f as the identity function for simplicity and defer formulations of more accurate fs to future research.

D. Parameters

Aligning the robot's motion and its speech requires us to be able to control their pace. In our controller, this is done by varying the **Audio Pace** a and the **Physical Pace** p.

Audio Pace (a) is the pace at which the robot's speech is played. E.g., $a = 1.2$ means the audio is played 20% faster than normal and $a = 0.8$ means playing 20% slower. We pass a to a *Phase Vocoder* [23] which time-scales the prerecorded speech audio with Short-time Fourier Transform. We empirically constrain $a \in (0.6, 1.4)$ to avoid incongruity arising from over-stretching or shrinking of the speech audio.

Physical Pace (p) is a variable in our admittance controller. It can be thought of as a "speed knob" with which we vary the controller. We design p to achieve the following effect: assuming constant F_{ext}, $p = 1.2$ should cause our controller to complete a trajectory in 20% less time than a fixed admittance controller following (2) and likewise $p = 0.8$ should cause it to run 20% slower. We empirically constrain $p \in (0.6, 1.4)$ so the admittance parameters do not deviate too much from their base values. *Physical Pace (p) must not be confused with end effector velocity. The role of p is not to directly modulate velocity, but to adjust the admittance parameters.* This ensures that the controller operates within the safety and compliance boundaries set by the admittance control framework. Directly multiplying the end effector velocity by p could lead to unsafe conditions, as it would bypass these regulatory mechanisms. Therefore, p should be understood as a rate constant that modifies the admittance parameters to indirectly influence end effector velocity, maintaining safety and compliance even under varying external forces. The formal definition of Physical Pace p is in V-E.

Estimated time-to-complete (ETC) for audio and speech We define alignment of motion and speech by relating the duration of the two. Therefore, to modulate the paces, our controller naturally needs to continuously estimate these durations, or, since the start time is known, when the motion and speech will end. Concretely, our controller computes the ETC for audio ($\hat{t}_a$) and motion ($\hat{t}_x$). $\hat{t}_a$ and $\hat{t}_x$ are computed [3] *under base pace* $a = p = 1$[4].

Computing Paces p and a We compute the ideal paces, p^* and a^*, from the following optimization:

$$\text{minimize } (p - p_{\text{natural}})^2 + (a - a_{\text{natural}})^2, \text{ s.t. } \frac{\hat{t}_a}{a} = f\Big(\frac{\hat{t}_x}{p}\Big), \tag{9}$$

where we set $p_{\text{natural}} = a_{\text{natural}} = 1$. $\hat{t}_x/p$ and $\hat{t}_a/a$ are the ETC for motion and audio considering the current pace, connected with the alignment model f. When f is the identity function, the solution to (9) is

$$p^* = \frac{s+1}{s^2+1}, \quad a^* = \frac{s^2+s}{s^2+1} \text{ where } s = \frac{\hat{t}_x}{\hat{t}_a}. \tag{10}$$

We further update p and a following the equation below:

$$\dot{p} = k_p(p^* - p), \quad \dot{a} = k_a(a^* - a). \tag{11}$$

This control equation makes both paces converge exponentially to their optimal values[5].

User Cooperation We define Cooperation ($0 < c < 1$) as

$$c(t) = 1 - \int_0^t \alpha^{t-\tau} \frac{\|F_{\text{ext}}\|}{\|F\|_{\max}} d\tau, \tag{12}$$

where $\|F\|_{\max}$ is the maximum magnitude of resisting force and α is the decay factor. In practice, we apply a deadband filter to F_{ext} first to filter out sensor noises and friction.

We then extend (11) to

$$\dot{p} = k_p(p^* - p), \quad \dot{a} = k_a(a^* - a) - k_c(1 - c). \tag{13}$$

This allows our controller to slow down the speech when the user does not cooperate (high resistance) and resume ideal speech pace when the user fully cooperates (low resistance). (13) achieves our second objective that the robot's motion and audio must be aligned under varying user cooperation.

E. Varying Admittance Model with Physical Pace p

We now give a formal definition of the Physical Pace p and integrate it into the virtual dynamics defined in (2). Let $v^*_{\text{ref}}(t)$ be the reference velocity generated by an admittance controller per (2) (without p), and let $v_{\text{ref}}(t)$ be the reference velocity from our controller (with pace p). Assuming fixed $(F_{\text{ext}} + F_{\text{virtual}})$, we want

$$v_{\text{ref}}(t) = p v^*_{\text{ref}}(pt) \tag{14}$$

which, after integrating both sides, implies that a controller with pace p would reach a reference position in $1/p$ the time of a controller without p. E.g., when $p = 2$, a controller without p would take twice amount the time to reach the same position

[3] $\hat{t}_a$ is computed as the sum of audio length on a path constructed by repeated use of (21) from the current vertex in the phrasing graph (minus the duration played for the current audio). $\hat{t}_x$ is computed assuming $F_{\text{ext}} = 0$ (i.e., fully cooperative user) with simulation of (7) at 500Hz.

[4] Both paces are defined relative to a base pace of 1. $a = 1$ means the audio is played at the recorded rate (free of distortions); $p = 1$ means the trajectory is being run with the default / intended admittance parameters. We consider 1 to be the most natural/ideal pace.

[5] We do not directly set $p = p^*$ and $a = a^*$, which could lead to abrupt change of pace if p^* and a^* deviate from current values too much.

as a controller with p. More generally, pace p varies with time. Let p_t be the pace at time t, we want

$$v_{\text{ref}}(t) = p_t v^*_{\text{ref}}(\Phi), \text{ where } \Phi = \int_0^t p_\tau \, d\tau. \tag{15}$$

Here time Φ generalizes pt for time-varying p in (14). The assumption of fixed $(F_{\text{ext}} + F_{\text{virtual}})$ in the definition is critical: *p generally is not a scale factor to the end effector velocity but instead modulates the admittance parameters.* This approach is safer as p effects the velocity only indirectly through the admittance model, which ensures the safety and compliance of our controller under varying forces.

We achieve (15) with the variable admittance model:

$$v_{\text{ref}} = A(p_t) \cdot (F_{\text{ext}} + F_{\text{virtual}}), \tag{16}$$

$$\text{where } A(p_t) = \frac{1}{\frac{1}{p_t^2} M_0 s + \frac{1}{p_t} D_0 - \frac{\dot{p}_t}{p_t^3} M_0}. \tag{17}$$

Proof: Differentiate (15) & multiply both sides by M_0,

$$M_0 \dot{v}_{\text{ref}}(t) = p_t^2 M_0 \dot{v}^*_{\text{ref}}(\Phi) + \dot{p}_t M_0 v^*_{\text{ref}}(\Phi). \tag{18}$$

Expanding $M_0 \dot{v}^*_{\text{ref}}(t)$ by (2),

$$M_0 \dot{v}_{\text{ref}}(t) = p_t^2 (F_{\text{ext}} + F_{\text{virtual}}) - p_t^2 D_0 v^*_{\text{ref}}(\Phi) + \dot{p}_t M_0 v^*_{\text{ref}}(\Phi).$$

Substitute $v^*_{\text{ref}}(\Phi) = v_{\text{ref}}(t)/p_t$ by (15),

$$M_0 \dot{v}_{\text{ref}}(t) = p_t^2 (F_{\text{ext}} + F_{\text{virtual}}) - (p_t D_0 - \frac{\dot{p}_t}{p_t} M_0) v_{\text{ref}}. \tag{19}$$

whose simplification then leads to (17). ■

E. Passivity Guarantees

We show that our controller defined in (17) is passive. A passive system is a system that is constrained in such a way that it does not inject excessive energy or instability into the interaction [24]. Formally, a system is passive w.r.t. an input-output pair $(u(t), y(t))$ if and only if there exists a positive definite storage function V over the system such that:

$$V(t) - V(0) \leq \int_0^t u(t)^T \cdot y(t) dt \quad \forall t > 0 \tag{20}$$

Theorem 1: *Consider a controller of the form outlined in* (17) *operating with linear trajectory T. If K is orthogonal and positive definite, D_0 is positive definite, and p is lower-bounded by a positive value, then the system is passive with respect to the force-velocity (F_{ext}, v_{ref}) input-output pair.*

(Proof provided in the Supplementary.)

[6]Results for all the users are available in our online appendix https://language-playback-robot-controller.github.io/user-sessions/.

[7]Step-like patterns for $\hat{t}_a$ around $t = 5$ and $t = 7$ is due to minor imperfection in our audio code when transitioning between different phrases at the time of the study. The steps should be lines of the same slope as the immediately preceding line and don't affect the correctness of our analysis.

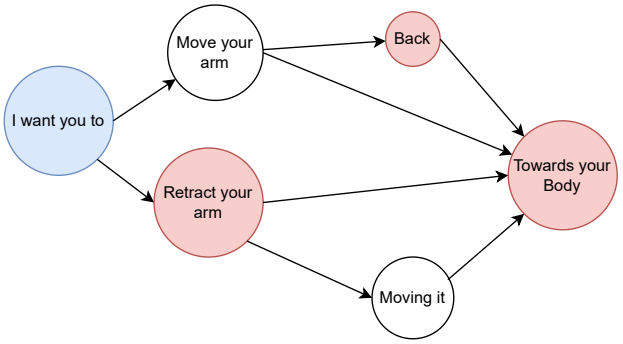

Fig. 3: Example phrasing graph for our experiments where users were asked to retract their arms. Paths ending at red nodes represent different phrasings of the instruction, e.g., "I want you to retract your arm" (short) or "I want you to move your arm back towards your body." (long)

F. Adaptively Paraphrasing the Robot Speech

To enable the robot to dynamically paraphrase its speech so the speech length matches that of the robot's motion (our third objective), we represent of speech as a **phrasing graph**. A phrasing graph is a Directed Acyclic Graph (DAG) where vertices denote sequences of words or phrases and a directed edge from vertex u to v denotes that the v's phrase could follow u's in speech. Phrasing graph captures the various alternative ways to express similar meanings (see Fig. 3).

When the controller finishes saying the phrase on a vertex, it chooses the next vertex/phrase based on how long it expects the trajectory to last. Formally, it does so by choosing a next vertex u from the graph, such that:

$$u = \underset{\text{next node } u}{\arg\min} \left| \hat{t}_x - \frac{\hat{t}_{\min}(u) + \hat{t}_{\max}(u)}{2} \right| \tag{21}$$

where $\hat{t}_x$ is the expected time-to-completion (ETC) of the trajectory (V-D). $t_{\min}(u), t_{\max}(u)$ are the minimum & maximum time-to-completion of speeches starting at vertex u[8]. We use the average of $t_{\min}(u)$ and $t_{\max}(u)$ as an efficient heuristic to increase the freedom of choices in subsequent vertex-selection steps: assuming the length of the speech starting at a node follows a balanced distribution, choosing the node with the closest min-max average to $\hat{t}_x$ reserves the flexibility for the algorithm rephrase in the future no matter in which direction $\hat{t}_x$ changes by then.

We remark that (a) (21) causes the controller to select longer paraphrases when the user resists more. Higher resistance slows the motion, increasing trajectory time $\hat{t}_x$. (b) the graph traversal does not depend on a or p to avoid compound effects[9].

[8]$t_{\min}(u)$ and $t_{\max}(u)$ are pre-computed by storing at each vertex the expected time to say its phrase, and iterating in reverse topological order.

[9]If we consider p here, i.e., we choose next vertex v that minimizes $|\hat{t}_x/p - \dots|$ instead of $|\hat{t}_x - \dots|$, a brief episode of low p induces a longer speech which requires even lower p to align – a vicious cycle.

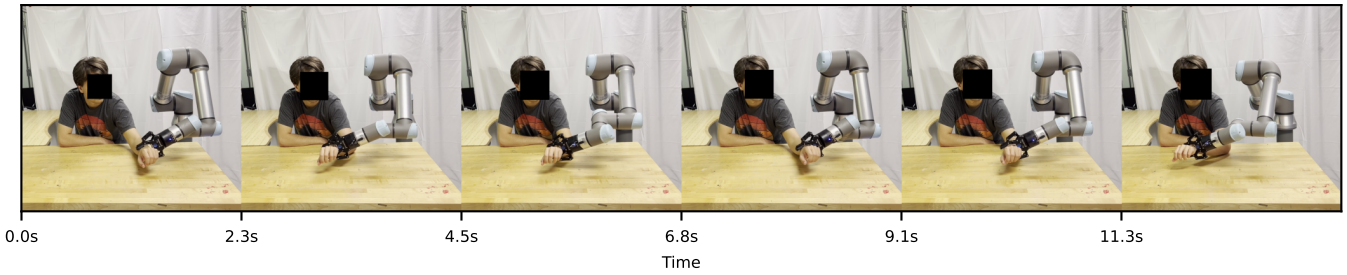

Fig. 4: A user participant is seen interacting with the UR5 robot on a desired trajectory, which was inspired by the therapy session ("shoulder external rotation") and predefined in our controller.[6] The reference trajectory for this motion was predefined. As evident in the figure, the user sits beside the robot and places their hand on the table surface, while the robot end-effector guides the user's arm through the planned trajectory.

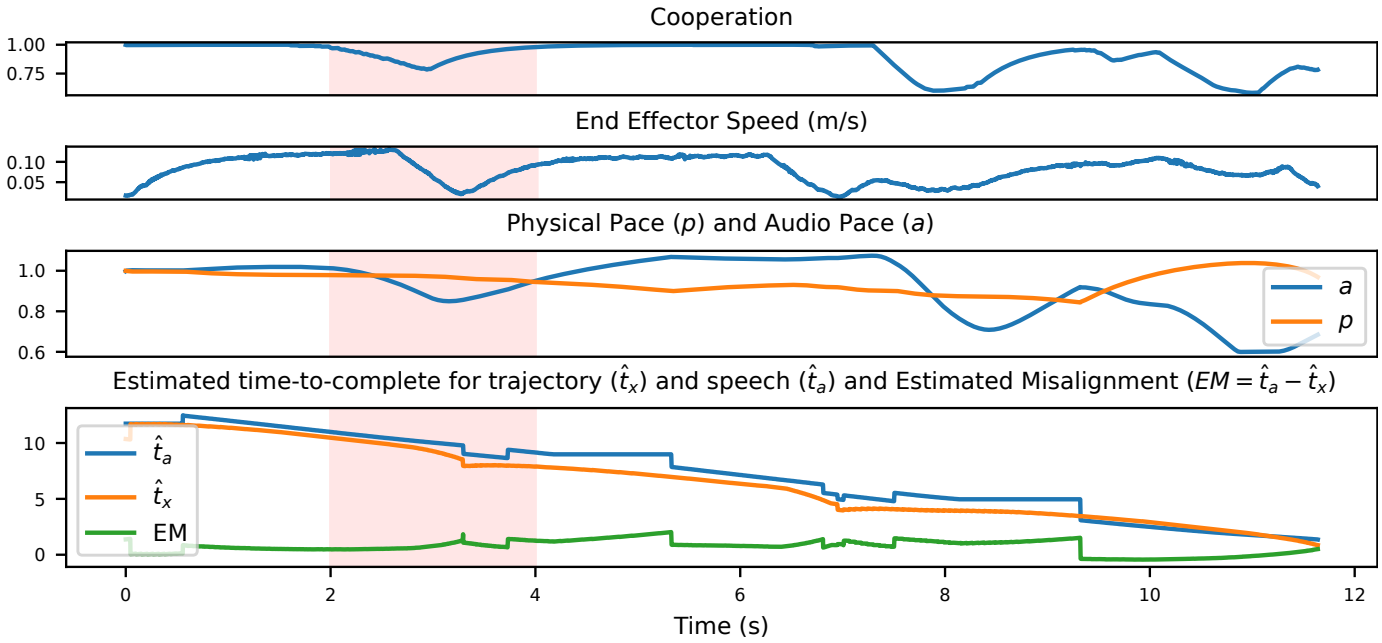

Fig. 5: States and metrics of our controller from a single user session through a predefined intended trajectory. At $t = 3$ (red shaded timespan), cooperation dropped, causing end effector speed to decrease. This shows that our motion controller is compliant. Audio pace also dropped and returned back up at $t = 4$, which shows our controller adapts the pace of the robot's speech to changing user cooperation. Throughout the session, our controller kept the estimated time-to-completion (ETC) of speech and trajectory close (the estimated misalignment is around 0). This demonstrates that our controller successfully aligns the robot's speech with its motion. Around $t = 7$, ETC for trajectory $\hat{t}_x$ dropped, making the controller select a shorter path on the phrasing graph, as shown by an abrupt drop in $\hat{t}_a$. However, the temporary dip in $\hat{t}_x$ was caused by inaccuracies of estimation, so the controller reverted to its original planned next phrase shortly after. As resistance dropped around $t = 9$, the controller eventually paraphrased and chose the shorter path. The paraphrase shortened the ETC to less than the trajectory, so the controller re-aligns by reducing audio pace and increasing physical pace. Overall, this demonstrates that our controller is able to align the robot's speech with its motion while adapting to changing user cooperation.[6] [7]

Note: the step-like pattern for $\hat{t}_a$ at around $t = 5$ and $t = 7$ are due to a minor imperfection in our audio player when transitioning between audio fragments of different phrases at the time of the study. The steps should be viewed as lines of the same slope as the immediate preceding line and does not affect the correctness of our analysis.

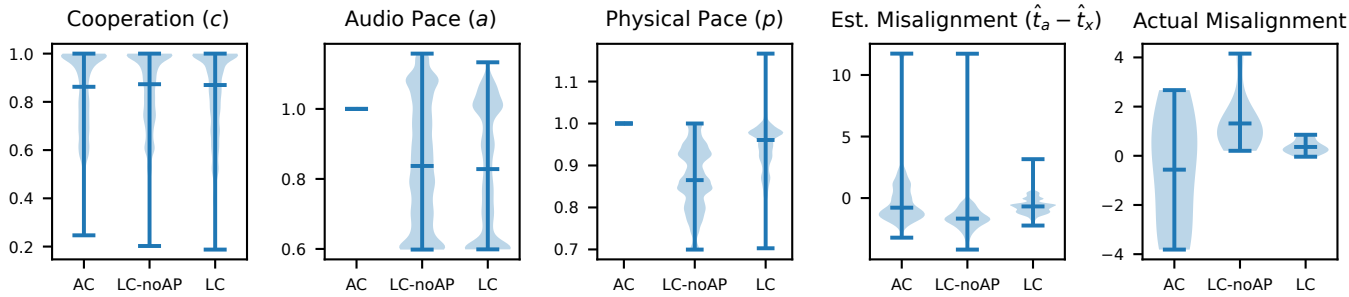

Fig. 6: Violin plot of the distributions of metrics for 3 controllers across all 12 user sessions. Blue bars indicate min, median, and max of a distribution. Consistent distributions of cooperation for all controllers show that all controllers deliver similar physical experience to the users; On this ground, LC exhibits less audio-motion misalignment (both actual and predicted) than LC-noAR which exhibits significantly less variation in misalignment than the AC baseline. Adaptive paraphrasing allows LC to maintain a more natural speed of speech most of the time than LC-noAR, as shown by a more concentrated peak around the default rate of 1 in audio rate distribution. Overall, LC best aligns the robot's speech with motion[6].

VI. EXPERIMENTATION

A. User Study

The user study included 12 participants (7 males, 5 females) with a mean age of 23 years. The study procedures and protocol were reviewed and approved by the Institutional Review Board (IRB Protocol #2212000845R001), ensuring we followed ethical guidelines for human subject research. The UR5 robot [1] guided users through a predefined trajectory demonstrated by the therapist. Users sat beside the robot, placed their hand on the desk. The robot guided their hand along the trajectory with speech instructions while users varied resistance arbitrarily.

B. Control Schemes Evaluated and Compared

1) **Admittance Controller with Decoupled Audio (`AC`)**: Our baseline is a pure admittance controller with dynamics of (2). The audio is a single prerecorded audio file that starts simultaneously with the trajectory. Physical and Audio Paces are not modulated at all.
2) **Language Controller without Adaptive Paraphrasing (`LC-noAP`)**: Our controller but without the ability to adaptively paraphrase the speech.
3) **Language Controller (`LC`)**: Our controller with adaptive paraphrasing by traversing a phrase graph (V-F).

C. Evaluation Metrics

1) **Cooperation** (c), defined in (12), quantifies user cooperation. c should not significantly vary between control schemes. In particular, our controller must not induce lower cooperation than the baseline.
2) **Audio Pace** (a) is the speed at which we deliver the speech audio. An audio rate of 1.2 indicates that we are playing the audio at 120% of its normal speed and 0.8 indicates playing at 80% of its normal speed. A pace closer to 1 implies less distortion from audio processing and more natural speech.
3) **Physical Pace** (p) is the state variable through which we vary the admittance parameters of our controller. p closer to 1 implies the robot operating closer to its most natural admittance parameters. Always 1 for AC.
4) **Actual Misalignment** is defined as the difference between the audio and motion durations, adjusted with the alignment model f. A positive value indicates the audio is longer than what would be ideal for the duration of the motion; A negative value indicates otherwise. A smaller absolute value suggests better audio-motion alignment.
5) **Estimated Misalignment** (EM $= \hat{t}_a - f(\hat{t}_x)$) is the real-time estimate of Actual Misalignment (AM). A smaller absolute value suggests better audio-motion alignment; EM differs from AM in that EM is a time series computed throughout the session whereas the AM is a scalar obtained after the session ends.

Of these metrics, Actual Misalignment is a single value per session, while the rest come in time series.

D. Analysis of Language Controller on a User Session

We analyze our controller's behavior for one of the user sessions in Fig. 5 and present a deep dive analysis. Around $t = 3$, cooperation dropped, slowing end effector speed. Audio pace dropped then recovered, aligning with motion. Throughout, estimated time to completion (ETC) of speech and trajectory were kept close, showing speech-motion alignment. Around $t = 7$, a temporary inaccurate ETC drop caused paraphrasing to a shorter phrase path, which was committed once resistance dropped around $t = 9$. We also observe a clear correlation between the audio rate a and the cooperation c, which proves that our controller slows down its speech to synchronize with the slower motion of the robot as we desired.

Overall, Language Controller adapts speech pace to follow motion changes from user cooperation variations, while paraphrasing aligns speech ETC to trajectory ETC.

E. Comparing Controllers across User Sessions

We compare all three control schemes across all users and present the results in Fig. 6. The level of user cooperation was similar across controllers, with no notable differences

observed initially. This implies that all users exhibited similar physical motion irrespective of controllers, supporting our prior hypothesis that alignment of robot's motion and language can be achieved while keeping users' physical interactions unchanged. AC exhibited broad misalignment (-4 to 3 seconds). Given short session length, this indicates notable lack of alignment. Adding pace control (LC-noAP) reduced misalignment, and adaptive paraphrasing (LC) further improved to under 1 second. Both LC finish speech after motion. LC controllers showed bimodal audio pace distribution - a peak around base rate 1, another at lower end – corresponding to two behaviors: slowing speech during high resistance to align with slower motion, and aiming for natural rate without/low resistance. LC has more concentrated peak at 1, as adaptive paraphrasing matches speech content length to motion, enabling natural rate delivery. We remark that the distribution of estimated misalignment for LC is irregular. We attribute this irregularity to adaptive paraphrasing, where the controller often updates the estimated time-to-completion for audio, $\hat{t}_a$, in a discrete manner owing to the naturally discrete distributions of sentence lengths on the phrasing graph

The experiments show that our controller met all of our defined objectives. It guided the users through the intended trajectory while adapting to the changing user resistance, and it controlled the pace and content of the robot's speech to maximize its alignment with the robot's motion.

VII. LIMITATIONS

We list below the limitations of our approach, which can provide insights to future investigations:

1) We employed variable admittance controller for motion control. But the idea generalizes to other controllers;
2) We proposed a general concept of speech-motion alignment where the durations of the two are linked with a monotonic function f. For simplicity, we set f as the identity function in this paper as it matches our empirical observations under the particular evaluation scenario. Further research should develop a more general and accurate alignment model.
3) We observe that in some of the therapy videos, the therapist varied the duration of her speech with the insertion of pauses and filler words. Our model does not implement this behavior, but in theory, it can be incorporated into our model by adding appropriate filler/pause nodes in our phrase digraph.
4) For richer interactions, large language models could automatically create phrasing graphs, beyond our therapist-recorded interactions;
5) Our controller assumes prerecorded audio phrases. An extension is using text-to-speech to generate audio from the phrases;
6) Formalized for physical therapy, Language Controller principles can be extend to manufacturing, with robots assisting in lifting, handling, and assembly; and space exploration, helping astronauts recover from falls.
7) Using TTS model also opens up more possibilities of speech modulation as many TTS models have configurable parameters such as speed and tone to control their generation.

VIII. CONCLUSION

We present a Language Controller to align a robot's physical motions and verbal utterances during collaborative tasks with humans. The controller was inspired by analyzing physical therapist interactions with patients. We identified key principles of adaptive pacing, aligned timing, and correlation between speech complexity and motion pace. These principles were formalized into control objectives to guide users through trajectories while modulating speech rate and content.

The Language Controller employed a variable admittance controller to enable compliant trajectory following. The pace of speech delivery was modulated via a phase vocoder to match the speed of motion and paraphrasing was incorporated through a phrase directed graph traversal in order to align the length of verbal content with the trajectory. Experiments with 12 users validated the advantages of our approach over baseline methods in synchronizing motion and speech.

Directions for future work include expanding the controller's language capabilities using large pre-trained models, applying it to diverse assistive tasks in areas like manufacturing, where robots work alongside humans in assembly; healthcare, where robots assist nurses with patient handling; and space exploration, where astronaut-robot teams assist from fall recovery and scientific explorations.

REFERENCES

[1] U. Robots, "Ur5 robot," https://www.universal-robots.com/products/ur5-robot/.

[2] R. Müller, M. Vette-Steinkamp, and K. Chatzilygeroudis, "Safe lifelong deployment of autonomous mobile robots in changing warehouses," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 234–243, 2022.

[3] K. Haninger, C. Viol, and K. Fischer, "Towards fluid human-robot handovers of arbitrary objects," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1197–1204, 2022.

[4] K. Barhydt and H. H. Asada, "A high-strength, highly-flexible robotic strap for harnessing, lifting, and transferring humans," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2110–2117, 2023.

[5] N. S. Selby, J. Ng, G. S. Stump, G. Westerman, C. Traweek, and H. H. Asada, "Teachbot: Towards teaching robotics fundamentals for human-robot collaboration at work," *Heliyon*, vol. 7, no. 7, 2021.

[6] T. Fong, C. Thorpe, and C. Baur, "Collaboration, dialogue, human-robot interaction," in *Robotics research: The tenth international symposium*. Springer, 2003, pp. 255–266.

[7] J. Thomason, S. Zhang, R. Mooney, and P. Stone, "Learning to interpret natural language commands through human-robot dialog," *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 1923–1929, 2015.

[8] D. Szafir, "The good, the bad, and the neutral: A meta-analytic review of human-robot collaboration," *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 49–58, 2021.

[9] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, 2011, pp. 1507–1514.

[10] C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox, "Learning language for robot control: Evaluating natural language understanding techniques," in *Experimental robotics*. Springer, 2013, pp. 439–452.

[11] J. He, J. Chen, X. He, J. Gao, L. Li, L. Deng, and M. Ostendorf, "Deep reinforcement learning with a natural language action space," *arXiv preprint arXiv:1511.04636*, 2015.

[12] J. Y. Chai, R. Fang, C. Liu, and L. She, "Collaborative language grounding toward situated human-robot dialogue," *AI Magazine*, vol. 37, no. 4, pp. 32–45, 2016.

[13] V. V. Unhelkar, S. Li, and J. A. Shah, "Decision-making for bidirectional communication in sequential human-robot collaborative tasks," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 329–341.

[14] F. Dimeas, N. Koenig, C. Savvaki, P. Asvestas, P. Sidiropoulos, E. Mavragani, N. Tsagarakis, and P. Artemiadis, "Online human-robot cooperation identification and adaptation to optimize robotic rehabilitation treatment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 11, pp. 1260–1270, 2016.

[15] L. Peternel, N. Tsagarakis, and A. Ajoudani, "Adaptive human-robot interaction via online action space adaptation," in *2017 International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2017, pp. 1340–1345.

[16] D. Gopinath, S. Jain, and B. D. Argall, "Human-aware manipulator path planning using iterative improvement optimal control synthesis," *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 2194–2201, 2017.

[17] C. Wang, C. Ross, Y.-L. Kuo, B. Katz, and A. Barbu, "Learning a natural-language to ltl executable semantic parser for grounded robotics," in *Conference on Robot Learning*. PMLR, 2021, pp. 1706–1718.

[18] Y.-L. Kuo, B. Katz, and A. Barbu, "Deep compositional robotic planners that follow natural language commands," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4906–4912.

[19] O. A. Can, P. Z. D. Martires, A. Persson, J. Gaal, A. Loutfi, L. De Raedt, D. Yuret, and A. Saffiotti, "Learning from implicit information in natural language instructions for robotic manipulations," *arXiv preprint arXiv:1904.13324*, 2019.

[20] Y. Bisk, D. Yuret, and D. Marcu, "Natural language communication with robots," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 751–761.

[21] A. Q. Keemink, H. van der Kooij, and A. H. Stienen, "Admittance control for physical human–robot interaction," *The International Journal of Robotics Research*, vol. 37, no. 11, pp. 1421–1444, 2018.

[22] N. Hogan, "Impedance control: An approach to manipulation," in *1984 American control conference*. IEEE, 1984, pp. 304–313.

[23] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[24] J. Wyatt, L. Chua, J. Gannett, I. Goknar, and D. Green, "Energy concepts in the state-space theory of nonlinear n-ports: Part i-passivity," *IEEE transactions on Circuits and Systems*, vol. 28, no. 1, pp. 48–61, 1981.