A Humanoid Robot Dialogue System Architecture Targeting Patient Interview Tasks

Yifan Shen*, Dingdong Liu*, Yejin Bang*, Ho Shu Chan, Rita Frieske, Hoo Choun Chung, Jay Nieles, Tianjia Zhang, Kien T. Pham, Wai Yi Rosita Cheng, Yini Fang, Qifeng Chen, Pascale Fung, Xiaojuan Ma, Bertram E. Shi

***Abstract*— Humanoid robots are promising approach to automating patient interviews routinely conducted by medical staff. Their human-like appearance enables them to use the full gamut of verbal and behavioral cues that are critical to a successful interview. On the other hand, anthropomorphism can induce expectations of human-level performance by the robot. Not meeting such expectations degrades the quality of interaction. Specifically, humans expect rich real-time interactions during speech exchange, such as backchanneling and barge-ins. The nature of the patient interview task differs from most other scenarios where task oriented dialogue systems have been used, as there is increased potential of engagement breakdown during interaction. We describe a dialogue system architecture that improves the performance of humanoid robots on the patient interview task. Our architecture adds a nested inner real-time control loop to improve the timeliness of the robot's responses based on the notion of "stance", an elaboration of the concept of a "turn", common in most existing dialogue systems. It also expands the dialogue state to monitor not only task progress, but also human engagement. Experiments using a humanoid robot running our proposed architecture reveal improved performance on interview tasks in terms of the perceived timeliness of responses and users' impressions of the system.**

I. INTRODUCTION

Medical staff routinely conduct patient interviews in a variety of settings, e.g., ward admission [1] (Fig. 1), cognitive assessment [2], etc. Automating some of these interview tasks with humanoid robots might help improve the efficiency of the healthcare system, by offloading repetitive and patient-independent portions to robots, enabling human staff to focus more on patient-specific concerns [3]. There are several potential advantages of humanoid robots for this task. Anthropomorphism makes people more willing to interact with humanoids [4]. It also facilitates the elicitation of pertinent information, since people generally regard humanoids as being sociable, trustworthy and empathetic [5], [6]. These are important social attributes for patient interviews [7], [8].

On the other hand, there are also challenges.

All authors are with Hong Kong University of Science and Technology in the Department of Electronic and Computer Engineering (YS, YB, HSC, RF, HCC, JN, YF, PF, BES), Department of Computer Science (DL, TZ, KTP, QC, XM), Center for Language Education (WYRC), and the Center for Aging Science (BES). This work was supported by the HKUST Center for Aging Science (projects FK081 and Z1011).

All supplementary materials can be found in `https://drive.google.com/drive/folders/1VyNDcMi_HWkoMJ-Ddp_eYz0tLZSHnHgI?usp=sharing`.

*These authors contributed equally to this work.

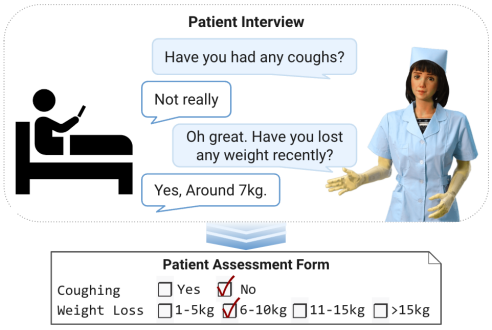


Fig. 1: An android interviewing a patient for ward admission.

First, anthropomorphism induces expectation of human-like performance [4]. Specifically, human's expect spoken dialogues to consist of a wide range of interactions beyond the simple exchange of words. Alongside speech exchange, people produce backchannel cues and coordinate spontaneous interjections (barge-ins). These auxiliary behaviors are essential to effective interviewing [9], [10], [11]. Not meeting these expectations leads to feelings of discomfort (the "Uncanny Valley" effect [12]) and lowers the overall quality of the interaction significantly [13].

Second, the nature of the patient interview task introduces new considerations not common in other task-oriented dialogues. Most task-oriented dialogue systems assume the human user is motivated to perform the task, e.g. seeking information. However, in the patient interview, the robot is seeking information from the human patient. Patients are not necessarily motivated or cooperative. Their state/mood may also change dynamically during the interview, rendering them unwilling or unable to continue. Hence, the robot needs to monitor patient engagement continually and take appropriate action (e.g., reach out for human assistance) if necessary.

Existing dialogue systems that have been applied to patient interviews cannot cope with the aforementioned challenges. Common architectures of task-oriented spoken dialogue systems (Fig. 2a) assume a turn-based pattern of alternations between the robot and human [14]. They also usually assume the patients to be cooperative and capable of answering interview questions. Thus, their dialogue state tracking often focuses solely on interview progress, with little consideration for monitoring patient engagement.

We describe a dialogue system architecture for humanoid robots targeting patient interview tasks (Fig. 2b), which can address the challenges identified in a formative study using a baseline system that implements the commonly-adopted "turn-based" dialogue system architecture. In contrast to the baseline architecture, our proposed architecture introduces the following innovations. First, we elaborate the binary concept of a "turn" to the more nuanced concept of a "stance" towards the conversational floor by the dialogue participants. Second, rather than a purely event-based feedback loop, where robot behaviors are determined at turn transitions, our architecture uses a nested feedback architecture, where decisions are made both at the event level (to control the overall dialogue/task state) and in real time (to control the robot's stance and behavior and the human's perception of the robot's stance). Third, we expand the dialogue state to track both task progress and patient engagement. Experiments with an humanoid robot using our proposed architecture show improved effectiveness in patient interview tasks compared to the baseline.

II. Formative Study Based on Prior Work

Past work automating medical interviews have all used a "standard" architecture for spoken task-oriented dialogue (ToD) systems (Fig. 2a), which follows a "turn-based" approach. Houser et al. used interactive voice response systems to make regular follow-up calls [15]. Virtual embodied conversational agents have been created to interview patients on alcohol consumption [16], post-traumatic stress disorder [17], etc. In [18] a humanoid interviewed the patient about diabetes self-management using tools such as quiz. In [19] humanoid robots were used as receptionists to collect patient information and give instructions upon admission. The humanoid robot Erica was used as a companion during COVID in [20].

A. Formative study design

To evaluate the performance of this standard architecture in patient interview tasks, we conducted a formative user enactment study with clinicians serving as proxies for patients, using their direct knowledge of patients gleaned from their clinical experience. Playing the roles of patients, the clinicians interacted with a humanoid robot [21] who interviewed patients following a script for the Abbreviated Mental Test (AMT) [22]. (See supplementary materials.)

Clinicians commented that quite frequently, patients become unwilling or unable to continue with interviews due to factors such as agitation or disorientation. To assess how the robot might perform under the range of patients encountered in clinical practice, we asked the clinicians to interact with the robot several times, role-playing different types of patients they commonly encountered.

After the interactions, we conducted semi-structured interviews where we reviewed video recordings of the interactions with the clinicians and solicited their comments on robot behaviors.

In this formative study, the robot was controlled by a baseline spoken dialogue system that follows the standard turn-based architecture (Fig. 2a). It consists of a real-time layer and an event-based layer [23], [14]. The real-time layer handles playback of a scripted sequence of behaviors and includes a turn monitor to identify transitions, i.e., end of human/system speech. The event-based layer is invoked upon turn transitions.

During the robot turn, a behavior executor executed a manually authored behavior sequence by controlling the servo motors and a Text-to-Speech (TTS) module. The behavior executor signalled The end of the robot turn directly to the turn monitor.

During the human turn, the robot listened to human speech and transcribed it using an Automatic Speech Recognition (ASR) module. The robot detected the end of the human turn by applying a duration threshold to periods of silence in ASR output.

After the human turn, transcriptions were interpreted via DialogFlow [24] using a rule-based dialogue manager (DM). Depending on task progress, the DM would choose an appropriate response, e.g., repeat the last question or ask the next question. Based on the DM's choice, the behavior generator selected one from a set of manually authored behavior sequences, which were passed to the behavior executor for execution during the robot turn.

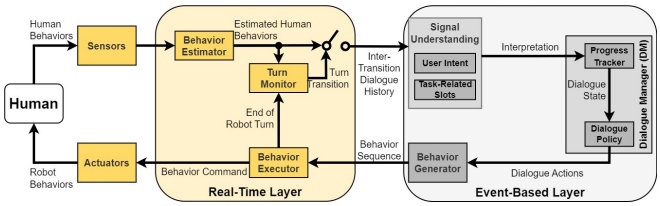
B. Results

The clinicians appreciated the robot's human-like appearance and facial expressions, which they believed could help patients feel more comfortable and engaged in the dialogue.

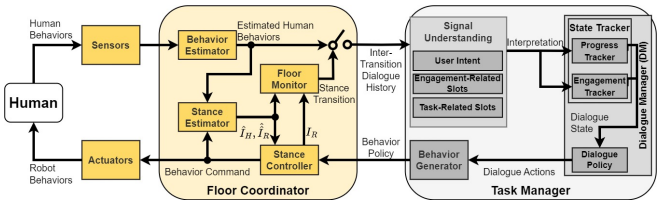
However, they pointed out that the timing of the robot's speech did not seem natural. Sometimes the robot did not respond to their utterances in time. On other occasions, the robot would abruptly speak when the clinicians did not expect, e.g., interrupting them while they were talking. The clinicians also complained that in cases where they sought to interrupt the robot (barge-in), e.g., to clarify a point or correct a previous mistake, the robot simply ignored them and continued with its ongoing behavior sequence. The clinicians commented that it would have been better for the robot to respond by stopping speaking and listening to the patient.

Interestingly, our findings suggest that the "Uncanny Valley" effect could have magnified the clinicians' reactions to these deficiencies. In the interviews, the clinicians mentioned a less human-like robot [25] already used in their hospital, which broadcasts notices and answers simple questions. Although its response delays and capability to handle interruptions were almost identical to those of our prototype system, the clinicians had few complaints, suggesting that they had different expectations and adjusted their behavior accordingly.

The clinicians also noted that during the interviews where they were role-playing disoriented or agitated patients, the robot simply persisted with asking interview questions despite receiving no, uncooperative or even hostile responses.



(a) Baseline Architecture



(b) Proposed Architecture

Fig. 2: Baseline (a) and proposed (b) architectures. Yellow/grey boxes are continuously-running/event-based.

Clinicians suggested that this behavior could be irritating and inappropriate.

Based on this formative study, we identified two main challenges, which our proposed architecture seeks to address.

1) Timeliness: Complaints about timeliness of the robot's responses stemmed from two sources. First, utterances by the robot in response to human speech appeared to be weirdly timed. After the human speech, the robot was silent and motionless as the behavior executor awaited instructions from the event-based layer, which was interpreting the human speech and formulating its response. This led to confusion about whether or not the robot had heard the human and about the robot's intention to speak. Humans occasionally spoke to follow up, leading the robot to interrupt them when it did respond. Second, the robot ignored human speech that occurred while it was speaking.

We traced these problems to deficiencies in the standard turn-taking model [26]. This model is binary, either one party is speaking or the other. It ignores other possible states of the conversational floor, e.g. where neither or both intend to speak. Timeliness issues arose due to misperceptions of the floor, both human misperception of robot intent and robot misperceptions of human intent.

Misperceptions of the first type can be avoided by backchannel cues. Human listeners backchannel to indicate that they are paying attention. Speakers backchannel to hold their turn while formulating speech [27], [28], [29]. While robot backchanneling has been studied, these behaviors are typically sequenced in the event-based layer by rules and/or data-driven methods [30], [31], [32], which leads to delays that degrade timeliness.

Misperceptions of the second type (robot ignoring human interruptions) arise because barge-ins are generally considered as exceptions to be avoided by the turn-taking model. While some studies did explore barge-in resolution via rules [33], [34] and/or data driven methods [35], [36], these were also handled by the event-based layer.

As we describe in the next section, we propose to handle these issues in an integrated manner at the real-time layer using a stance-based framework, which models the state of the floor as combinations of stances (intent to speak or listen) by the two dialogue participants (robot and human). While stance-based frameworks have been proposed previously [37], problems arising from misperceptions of the floor and their real-time resolutions have not been considered. Our framework also enables context-dependent handling of barge-ins similar to the way humans do [34].

2) Engagement Breakdown: Problems with engagement breakdown arose because the baseline architecture solely tracks task progress. Most dialogue systems proposed previously for healthcare lack considerations of usability and engagement [38]. We propose to enrich signal understanding [39] and track an expanded dialogue state that encompasses both task progress and patient engagement. This facilitates the recognition and handling of engagement breakdown.

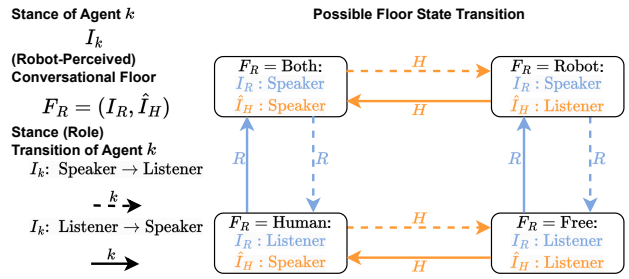
III. Proposed Architecture

A. Overview

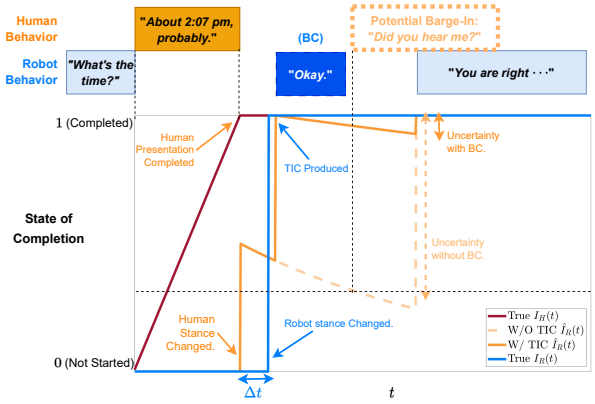
The proposed architecture is illustrated in Fig. 2b. A humanoid robot provides a stream of sensory input, such as audio and visual (image) steams. They are sent to the continuously-running Floor Coordinator (FC), where a behavior estimator estimates human behaviors, e.g., speeches, facial expressions, gestures, etc. Using both estimates of human behavior as well as knowledge of the robot's own behavior, a stance estimator estimates the human stance and human perception of robot stance. Stance estimates are fed back to the stance controller (StC), which takes the place of the behavior executor in the baseline architecture, so that the robot can adjust its behavior (utterances, facial expressions and gesticulations) in real time to manage the human stance and the human perception of the robot stance. Unlike the turn-based baseline architectures, the FC adds a continuously-running feedback pathway via the stance estimator and the stance controller, which is critical in improving timeliness. Stance estimates are also passed to the floor monitor, which checks for floor state transitions, which serve as a unified representation of both normal turn transitions and barge-ins (Section III-B.1).

At a floor state transition, the FC invokes the Task Manager (TM) and passes it the dialogue history which encodes the multi-modal behaviors from both parties during the last inter-transition period. The structure of the TM is similar to that of the event-based layer in the baseline architecture, but the individual components possess enhanced capacities. The signal understanding component (SgU) interprets the inter-transition history for dialogue intent and semantic entities that are related to task progress and/or engagement level (Section III-C.1). The dialogue manager (DM) tracks an expanded dialogue state encompassing both task progress and engagement level (Section III-C.2).

Behavior generation maps the dialogue action to a behavior policy, rather than a scripted sequence, which is sent to the stance controller in the FC.



(a)



(b)

Fig. 3: Entities pertaining to human/robot are orange/blue colored. (a) illustrates stance and conversational floor. (b) is a robot stance model. The blue curve is the actual robot stance. The solid and dashed orange curve represent human-perceived robot stance, w/ and w/o backchannel cues (BC).

B. Floor Coordinator

1) Stance-Based Formulation: The FC adopts a *stance*-based formulation (Fig. 3a). We index the robot/human agent with $k \in \{R, H\}$. Agent k's "stance" at time t, $I_k(t)$, has two components: a binary component ("Speaker"/"Listener") and a continuous component representing the agent's progress in realizing its intent (i.e, the progress of speech delivery/recognition). Perceptions of stance by the other agent is indicated by a "hat", e.g., the robot-perceived human stance is $\hat{I}_H(t)$. The configuration of the two interlocutors' roles constitutes the state of the conversational *floor* [37]. $F_R(t) = (I_R(t), \hat{I}_H(t))$ is the robot-perceived floor. We index stance (role) transitions by $i \in \mathbb{N}$. We denote the inter-transition period as $T_i = [t_i, t_{i+1})$. Fig. 3a illustrates transitions between possible floor states.

The stance-based formulation is more general than the turn-taking model. A robot-perceived turn corresponds to an inter-transition period T_i where $F_R(t)$ = "Human" or "Robot". Turn transitions correspond to floor state transitions where the two agents switch roles (consecutively). The person yields the turn if (estimate of) human stance changes from speaker to listener ($\hat{I}_H(t_i)$: Speaker → Listener) while $F_R(t)$ has been "Human". Upon termination of robot speech, the robot changes the binary component of I_R from speaker to listener. Barge-ins are also represented as stance transitions. The robot perceives a human barges-in if $\hat{I}_H(t_i)$: Listener → Speaker while $I_R(t_i)$ = Speaker.

Human stance is estimated by a human stance model to produce $\hat{I}_H(t)$. A silence threshold applied to a voice activity detector (VAD) can estimate the binary component of human stance (role). A floor monitor detects transitions in $F_R(t)$ and invokes the TM. Since barge-ins are also represented as transitions in $F_R(t)$, the TM processes them via the same pipeline like normal turn transitions, thereby enabling context and policy based barge-in handling.

2) Stance Controller: The stance controller (StC) coordinates the choice and timing of robot behaviors within T_i. These include robot stance transitions and behaviors such as speech delivery and backchanneling. The StC's coordination is configured by the behavior policy P_i. P_i could be a behavior sequence similar to the standard architecture, an explicit policy, an implicit policy specified by a cost function, or a combination of those.

The idea of specifying P_i as a cost function emerges from the stance-based formulation and provides a functional motive for backchanneling. As discussed in Section II-B.1, a period of silence after human speech can lead to uncertainty in the human's perception of robot stance. The StC treats robot behaviors including backchanneling as means to control human perceptions of robot stance. The person is assumed to estimate robot stance ($\hat{I}_R(t)$). The StC estimates this human perception as $\hat{\hat{I}}_R(t)$ with a *robot stance model*.

If P_i is specified as a cost function punishing differences between $\hat{I}_R$ and I_R, the StC can use Model Predicative Control (MPC) [40] to derive robot actions. Given $I_R(t)$ over T_i, an alignment optimizer solves for a sequence of robot behaviors that minimizes the difference between $\hat{\hat{I}}_R(t)$ and $I_R(t)$ over a prediction horizon.

Fig. 3b illustrates an example of a robot stance model. The person finishes speaking (realizes his/her intention as a speaker) when the red curve reaches 1. Based on his/her behavior, s/he also expects that robot to realize s/he has finished speaking, and that the robot's role as a listener has ended, with some confidence (the orange curve goes to 0.5). After a silence threshold Δt, the robot actually detects the end of human speech (blue curve goes up to 1). If the robot remains inactive following the end of human speech, the person becomes more uncertain about the robot stance (the orange curve begins to decay). This could eventually lead the human to probe the robot stance (the dashed orange curve drops below the dashed black line). On the other hand, a backchannel cue could reduce uncertainty and align human perception with I_R (the solid orange curve is brought closer to the blue curve after the robot utters "Okay") Applying MPC to align $\hat{\hat{I}}_R(t)$ with $I_R(t)$ assuming this robot stance model enables the StC to derive and execute backchannel cues.

C. Task Manager

In typical ToD scenarios like hotel booking, users are self-motivated to engage with the system, as they have

Fig. 4: Illustration of signal understanding to obtain user intent and task-related and engagement-related slots.

specific objectives. Hence, the system only needs to track task progress, In contrast, patient interviews involve robot-initiated dialogues where user engagement cannot be presumed. This necessitates the estimation and management of user engagement.

1) Signal Understanding: The signal understanding (SgU) component processes the input signal S, the inter-transition dialogue history including floor state transitions and the two parties' behaviors. It produces interpretations in the form of the user's dialogue *intent* ($Int(S)$) and relevant *slots* filled with semantic concepts ($Slot(S)$). Intent detection determines the patient's aim. Slot filling identifies and assigns values v to key semantic entities k, yielding results of the form $\{k : v\}$.

As illustrated in Fig. 4, the semantic concepts underlying $Int(S)$ and $Slot(S)$ are related to task progress and the user's engagement level during the inter-transition period. Possible intents detected include `answer_question`, and `ask_for_help`, `not_answer_question`. Task slots generally hold patient answers to interview questions. Engagement slots reflect the user's willingness and ability to proceed.

Intent detection and slot filling were performed via large language models (LLM) in an instruction-based manner. For example, the prompt used for Fig. 4 is "*Evaluate how much the patient lost weight. Choose one of the categories: 1-5kg, 6-10kg, 11-15kg, >15kg, Unsure. If the patient does not answer the question, choose UNCLEAR.* [*Dialogue*] *Nurse:* {*robot_question*} *Patient:* {*human_response*}", where $robot\_question$ and {human_response} are filled according to the dialogue history. Since the answer slot is filled in this example, the user's dialogue intent was recognized as `answer_question` and the engagement slot `answer_intelligibly` is filled as *yes*. Another prompt issued in parallel was used to fill the `answer_willingly` slot: *You are a robot nurse who is conducting a patient assessment form survey in Cantonese. Is the patient being uncooperative? Answer yes only if the patient refuses to answer and is upset and curses or is afraid of you and wants to speak to a human nurse. Choose between two options: Yes, No.* More detailed examples are provided in "SgU_Examples" of supplementary materials.

2) Dialogue Manager: Based on interpretations from the SgU, the state tracker in the dialogue manager (DM) tracks the dialogue state, i.e., task progress and human engagement level. Future dialogue actions (e.g., asking question, seeking human interventions, etc.) are then decided by the dialogue policy.

In the standard architecture, the dialogue state developed from task-related interpretations represents task progress only (progress tracker). In contrast, the dialogue state of our DM is expanded to cover both task progress and the interlocutor's engagement (engagement tracker). The latter is supported by the engagement-related slots filled by the SgU. For instance, many `answer_intelligibly` slots flagged as `no` may indicate confusion in the user. Many negatively filled `answer_willingly` slots and/or the missing `answer_question` intent may reflect unwillingness to engage.

When the person appears engaged, the DM focuses on completing the interview task. In contrast, if the dialogue state indicates an engagement breakdown, the DM instructs the robot to disengage from the conversation and seek human intervention.

Notably, the engagement-related slots characterize the nature of the engagement breakdown, which provides additional context for subsequent human follow-up. The processing of SgU is parallelized. After one floor state transition, several dialogue state updates are made and multiple dialogue actions produced. Usually the DM first decides the next stance, and then more specific actions like what questions to ask.

3) Behavior Generation: Dialogue actions are mapped to behavior policies during behavior generation and sent to the StC (Section III-B.2). The behavior policy could take various forms. For example, it can be a combination of a cost function punishing human misperception of I_R, which leads to backchaneling, as discussed previously, and a behavior sequence. The dialogue action of repeating the current question is mapped to a paraphrasing of that question. To disengage from the conversation, the robot physically moves to a human nurse while saying either *"Sorry, this patient seems a bit upset, could you please come and help?"* or *"sorry, I am having issues with communicating with this patient, could you please come and help?"* depending on the patient's state of engagement, as discussed previously.

IV. Experiments

A. Coordination of Robot Actions

We evaluated the effectiveness of FC in coordinating robot behaviors including backchanneling and barge-in resolutions.

1) System Implementation: We used the same prototype system employed in our formative study for baseline. For experimental system, we replaced the prototype's real-time layer with an initial implementation of the FC module and made minor modifications on the event-based layer for compatibility. The "AMT_demo" in supplementary materials is a video of the experimental system conducting the AMT. Following the model depicted in Section III-B.2, the FC backchanneled in robot turn while the event-based layer was processing ("bc_demo1" and "bc_demo2" in the

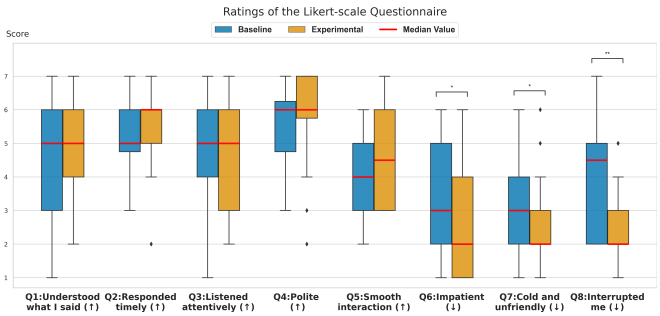


Fig. 5: Box plot of the ratings of the Likert-scale questions from Section IV-A. Red bar represents median. Diamonds are outliers w.r.t. this range. Significant pairwise differences were annotated above the boxes: '$*$' for $0.01 \leq p < 0.05$, '$**$' for $0.001 \leq p < 0.01$, and '$***$' for $p < 0.001$. Color differentiates system types. The bracketed arrow pointing up/down denotes that a higher/lower value is better.

supplementary materials are examples of these backchannel cues.). It also signaled human barge-ins as state transitions to the event-based layer, who would configure the FC to stop ongoing delivery and switch to listening stance.

2) Experimental Procedure: Twenty university students were recruited for this experiment. They participated in the same AMT task of the formative study twice, one with each system. After each interaction, we obtained the participant's subjective evaluation of the robot's behaviors via 7-point Likert-scale questionnaires as well as semi-structured interviews. In the supplementary materials, see "LikertQuestions" for the Likert-scale questionnaire and "OpenQuestions" for questions asked during this semi-structured interview. The order of interactions was randomized and counterbalanced.

3) Results: Fig. 5 illustrates results of the Likert-scale questions. Higher scores in Q2, Q5 and lower scores in Q8 indicate that the perceived timeliness of the robot was better (Cronbach's Alpha: 0.657). Higher scores in Q1, Q3, Q4 and lower scores in Q6, Q7 imply better impressions of the robot. A Wilcoxon signed rank test revealed that the experimental system (orange) was significantly better than the baseline system (blue) in Q8. The p-value p, test statistics W, effect size r (matched pairs rank-biserial correlation), and interpretation were ($p = 0.004, W = 106.5, r = 0.77$, large effect). While the trends for Q2 and Q5 were not statistically significant, they favored the experiment system. Among the rest of the questions, the experiment system was significantly better on Q6 and Q7, with ($p = 0.037, W = 61.50, r = 0.58$, large effect) and ($p = 0.035, W = 53.0, r = 0.61$, large effect) for Q6 and Q7 respectively. The trends for the other questions also favored the experiment system. Overall, these results suggested that the FC improved timeliness and was effective in coordinating backchanneling and supporting barge-in handling.

B. Signal Understanding of Task Manager

We further evaluated the accuracy of the interpretations from the signal understanding component. We focused on text input and implemented the SgU with GPT-4 for slot filling from texts. All test samples were derived from the interview task of filling the patient assessment form Patient Assessment Form (PAF) (Fig. 1). See "PAF" in the supplementary materials for the assessment form. Each sample included the robot's question, the user's reply, and the correct slot values. The transcriptions were all in Cantonese.

1) Task-Related Interpretations: This experiment evaluated the accuracy of task-related slots filled by our SgU. We compiled a test set of 86 samples collected from both synthetic and real interactions. No engagement breakdown was present in this dataset. The SgU achieved an accuracy of 95.35% in slot-filling, evidencing its effectiveness in producing task-related interpretations.

2) Engagement-Related Interpretations: This experiment explored the SgU's ability in engagement-related slot filling. We prepared another dataset consisting of 30 samples of actual interactions. 20 of them were interactions where engagement breakdown occurred due to the user being unwilling to continue, characterized by the presence of foul languages, among other indicators, in the user's response. For these user responses, the engagement-related slot `answer_willingly` should be filled with `No`. The remainder of the 10 interactions saw no sign of engagement breakdown but the user's response was assertive and carried intense emotions. Some of the (translated) responses are "*I'm coughing to death*" and "*Why are you asking me this, I haven't been coughing*". These samples examined whether the SgU would raise false alarms for edge cases. Results revealed that the slot was correctly filled in 28 out of 30 samples, leading to a 93.33% accuracy rate.

C. Overall Evaluation

Finally, we made a holistic evaluation of the proposed architecture through a user enactment study.

1) System Implementation: The experiment system was an implementation of the proposed architecture deployed on the same android of the formative study. The FC and SgU were the same as in Sections IV-A and IV-B. The DM was rule-based, and in behavior generation behaviors were chosen from manually-authored scripts. The baseline system used the same TM of the experimental system. But instead of the FC, it adopted the real-time layer of the prototype system in Section II. Minor modifications were made for compatibility.

2) Experimental Procedure: Four clinicians were recruited. They interacted with the android administering the same PAF task as in Section IV-B (Fig. 1). See "PAF_demo" for a video of the experimental system performing the PAF. Each subject underwent two experiment sessions. In the first session, s/he interacted with one of the two systems, i.e., baseline or experimental system. In the second session the other system was used. Each session had four interactions. In the first two interactions, s/he enacted the role of either a lucid and cooperative patient or a patient who was hard to work with. In the second two, s/he role played the other type of patient. Hence each subject had eight interactions in total. After each interaction, the subject filled the same Likert-scale

questionnaire used in Section IV-A. We also took notes of their comments regarding that interaction. A semi-structured interview similar to that of Section IV-A was conducted at the end of each session, where the subject's overall impressions and suggestions were discussed. The order of system and patient type was randomized and counterbalanced. In each interaction only a subset of questions from the PAF were asked. To prevent learning effect, we created 8 different question sequences. These sequences were of similar complexity. See "PAF_Sequence" in the supplementary materials for more details. They were used for all subjects, but the order with which each subject met them was randomized.

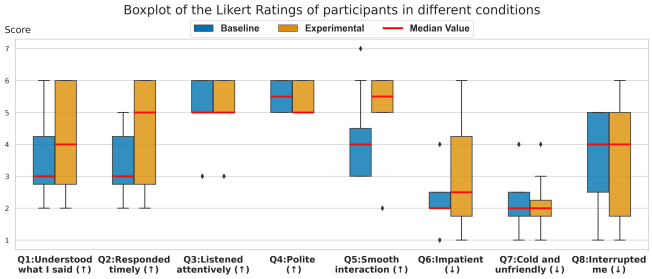
3) Results: When a cooperative and lucid patient was role-played (Fig. 6a), the ratings for Q2, Q5 and Q8 favored the experimental system over the baseline system. This supported the improved timeliness and responsiveness afforded by the FC of the experimental system. In interactions with difficult patients (Fig. 6b), the experimental system was rated worse than the baseline system on Q2, Q5 and Q8 as well as on several other questions. Due to the limited number of subjects, we can only hypothesize the cause here. Some clinicians noted that the robot's backchanneling *"sounded like the robot didn't hear me"* and the robot should adjust its behaviors *"if the person appeared impatient"*. I.e., backchannel cues produced by the FC could have further irritated patients with low level of engagement. It could be that robot's backchanneling should vary by human engagement level.

Of the 90 task-related slots occurred across all interactions with both system, 84.44% were correctly filled by the SgU. This result testified the efficacy of our SgU in making task-related interpretations of human response. Clinicians also appreciated the TM's capability of monitoring engagement and handling breakdown. For example, in one interaction a clinician who pretended to be impatient and uncooperative commented that the robot "*did it well*," when it recognized the patient's unwillingness and sought human intervention. (See "disengage_demo" in the supplementary materials for a video recording of this interaction.) Moreover, many clinicians commented that, now that the robot could perceive the patient's engagement level, it should incorporate interview techniques commonly adopted by clinicians.

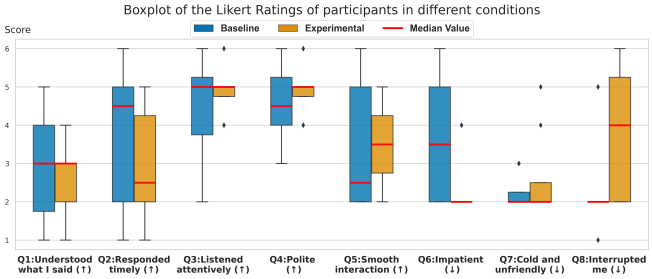
Overall, these results suggested that our proposed architecture improved the robot's effectiveness in administering the PAF task.

V. Conclusion

In this work, we proposed a novel dialogue system architecture for humanoid robots performing patient interview tasks. A floor coordinator coordinates robot behaviors including backchanneling and supports barge-in handling. Its coordination derives from the management of participants' stance and perception of stance. A task manager interprets dialogue history to track an expanded dialogue state encompassing both interview progress and patient engagement. Together, our architecture can better cope with many challenges arising from performing patient interviews with humanoid robots.



(a) Results when role-playing a cooperative and lucid patient.



(b) Results when role-playing a hard-to-work-with patient.

Fig. 6: Box plot of the Likert-scale questionnaire from Section IV-C. The convention of Fig. 5 was followed.

We identified several limitations and future directions. In terms of the experiments, our subject population is small and doesn't match that of the target population. Our implementation of our proposed architecture can also be improved. Future work should enrich the human/robot stance model by fusing the multi-modal behaviors of both parties. Likewise, currently the dialogue manager navigated through the interview task via simple rules. As suggested by the clinicians, the capability to track engagement should enable the robot to support more flexible dialogue policies such as asking scaffolding questions [41] if the person has difficulty answering. One interesting direction to explore would be LLM-based online generation of scaffolding questions. Notably, to meet the standards of medical setting, it will be paramount to keep the content of the dialogue safe and under control.

References

[1] M. L. Vancott, "Communicative Competence during Nursing Admission Interviews of Elderly Patients in Acute Care Settings," pp. 184–208, 1993.

[2] G. M. Giordano *et al.*, "The Cognitive Assessment Interview (CAI): Association with neuropsychological scores and real-life functioning in a large sample of Italian subjects with schizophrenia," *Schizophrenia Research*, vol. 241, pp. 161–170, 3 2022.

[3] T. Bickmore and T. Giorgino, "Health dialog systems for patients and consumers," vol. 39, pp. 556–571, 2006.

[4] R. Y. Kim, "Anthropomorphism and Human-Robot Interaction," *Communications of the ACM*, vol. 67, pp. 80–85, 2024.

[5] E. Broadbent *et al.*, "Robots with Display Screens: A Robot with a More Humanlike Face Display Is Perceived To Have More Mind and a Better Personality," *PLOS ONE*, vol. 8, p. e72589, 8 2013.

[6] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 33–42, 3 2020.

[7] K. Holmes and N. Turville, "Interview Techniques," *Developing Healthcare Skills Through Simulation*, pp. 231–242, 4 2023.

[8] C. D. Lauster and S. B. Srivastava, "Patient Interview," *Fundamental Skills for Patient Care in Pharmacy Practice*, pp. 1–346, 2014.

[9] M. Johansson *et al.*, "Making turn-taking decisions for an active listening robot for memory training," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9979 LNAI. Springer Verlag, 2016, pp. 940–949.

[10] D. Lala *et al.*, "Attentive listening system with backchanneling, response generation and flexible turn-taking," pp. 127–136.

[11] T. Yokozuka *et al.*, "The Relationship Between Turn-taking, Vocal Pitch Synchrony, and Rapport in Creative Problem-Solving Communication," *Speech Communication*, vol. 129, pp. 33–40, 5 2021.

[12] M. Mori, "The Uncanny Valley: The Original Essay by Masahiro Mori," *IEEE Robotics & Automation Magazine*, vol. 12, pp. 1–6, 2012.

[13] A. Sciutti and G. Sandini, "Interacting with robots to investigate the bases of social interaction," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, pp. 2295–2304, 12 2017.

[14] K. Jokinen and M. McTear, "Spoken Dialogue Systems," 2010.

[15] S. H. Houser *et al.*, "Telephone Follow-Up in Primary Care: Can Interactive Voice Response Calls Work?" *Studies in health technology and informatics*, vol. 192, p. 112, 2013.

[16] R. Amini *et al.*, "On-demand virtual health counselor for delivering behavior-change health interventions," *Proceedings - 2013 IEEE International Conference on Healthcare Informatics, ICHI 2013*, pp. 46–55, 2013.

[17] G. Stratou *et al.*, "A demonstration of the perception system in SimSensei, a virtual human application for healthcare interviews," *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, pp. 787–789, 12 2015.

[18] M. A. Neerincx *et al.*, "Socio-cognitive engineering of a robotic partner for child's diabetes self-management," *Frontiers in Robotics and AI*, vol. 6, 2019.

[19] H. S. Ahn *et al.*, "Hospital Receptionist Robot v2: Design for Enhancing Verbal Interaction with Social Skills," *2019 28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019*, pp. 1–6, 2019.

[20] E. Ishii *et al.*, "ERICA: An empathetic android companion for covid-19 quarantine," in *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, H. Li *et al.*, Eds. Singapore and Online: Association for Computational Linguistics, Jul. 2021, pp. 257–260.

[21] "Awakening Health – Awakening Health Through Robotics."

[22] H. M. Hodkinson, "Evaluation of a mental test score for assessment of mental impairment in the elderly," *Age and ageing*, vol. 1, pp. 233–238, 11 1972.

[23] L. Jacqmin *et al.*, ""do you follow me?": A survey of recent approaches in dialogue state tracking," in *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2022, pp. 336–350.

[24] "Dialogflow — Google Cloud."

[25] "temi - Robots as a Service - AI in Motion."

[26] G. Skantze, "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review," *Computer Speech and Language*, vol. 67, p. 101178, 2021.

[27] E. A. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in Society*, vol. 29, pp. 1–63, 2000.

[28] S. Duncan and G. Niederehe, "On signalling that it's your turn to speak," *Journal of Experimental Social Psychology*, vol. 10, pp. 234–247, 5 1974.

[29] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology*, vol. 23, pp. 283–292, 8 1972.

[30] G. Skantze and A. Hjalmarsson, "Towards incremental speech generation in conversational systems," *Computer Speech & Language*, vol. 27, pp. 243–262, 1 2013.

[31] G. Skantze *et al.*, "Exploring turn-taking cues in multi-party human-robot discussions about objects," *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, pp. 67–74, 2015.

[32] T. E. Lin *et al.*, "Duplex Conversation: Towards Human-like Interaction in Spoken Dialogue Systems," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3299–3308, 2022.

[33] N. Ström and S. Seneff, "Intelligent barge-in in conversational systems," *6th International Conference on Spoken Language Processing, ICSLP 2000*, pp. 1–4, 2000.

[34] F. Gervits and M. Scheutz, "Towards a conversation-analytic taxonomy of speech overlap," *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pp. 4598–4602, 2019.

[35] D. Lala *et al.*, "Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues," in *ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction*. Association for Computing Machinery, Inc, 10 2019, pp. 226–234.

[36] D. Bekal *et al.*, "Contextual Acoustic Barge-In Classification for Spoken Dialog Systems," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022-September. International Speech Communication Association, 2022, pp. 1091–1095.

[37] A. Raux and M. Eskenazi, "Optimizing the turn-taking behavior of task-oriented spoken dialog systems," *ACM Transactions on Speech and Language Processing*, vol. 9, 2012.

[38] M. Valizadeh and N. Parde, "The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan *et al.*, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6638–6660.

[39] Z. Wei *et al.*, "Task-oriented dialogue system for automatic diagnosis," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 201–207.

[40] M. Morari *et al.*, "Model predictive control: Theory and practice," *IFAC Proceedings Volumes*, vol. 21, pp. 1–12, 6 1988.

[41] P. A. Saunders *et al.*, "'Oh he was forgettable': Construction of self identity through use of communicative coping behaviors in the discourse of persons with cognitive impairment," *Dementia*, vol. 10, pp. 341–359, Aug. 2011.