Speech Perception and Implementation in a Virtual Medical Assistant

Aryana Collins Jackson[a], Yann Glémarec[b], Elisabetta Bevacqua[c], Pierre De Loor
and Ronan Querrec
ENIB, Lab-STICC UMR 6285 CNRS, Brest, France

Keywords: Situational Leadership, Medicine, Speech, Virtual Agent, Embodied Conversational Agent.

Abstract: In emergency medical procedures, positive and trusting interactions between followers and leaders are imperative. That interaction is even more important when a virtual agent assumes the leader role and a human assumes the follower role. In order to manage the human-computer interaction, situational leadership is employed to match the human follower to an appropriate leadership style embodied by the agent. Situational leadership was used to create 33 utterances indicative of the four different leadership styles. A participant evaluation was then carried out in order to examine (1) whether perceptions of leader trust and motivation vary dependent on both readiness level and utterance syntax and (2) whether follower ability and willingness are affected by the leader's speech. We found that general perceptions of leadership behavior influenced follower performance and that the leader's speech influences followers' ability. Finally, we demonstrate how the results of this study are implemented in a virtual agent system.

1 INTRODUCTION

The field of embodied conversational agents (ECAs) largely involves creating virtual agents that can communicate effectively with humans. The end goal is often to create an agent that converses and behaves like a human in order to complete a task. Previous work has devised virtual ECAs for companionship (Montenegro et al., 2019), for diagnostics (Philip et al., 2020), and for liaising with expert staff (Bickmore et al., 2015). However, sometimes a virtual agent must accomplish a goal that involves more than face-to-face communication. In cases such as these, the end goal may involve tasks that require human-agent cooperation (Lala et al., 2015; Ramchurn et al., 2015).

One domain that involves cooperation regardless of whether agents or humans are included is medicine. In a variety of tasks, cooperation is key to ensuring efficiency and patient safety (Araszewski et al., 2014). In emergency medicine, that cooperation is especially important as the patient is in a precarious state. Emergency medicine is complicated even further if the caregivers are not professional medical providers. Such a situation might occur during a remote expedition in which the humans present are not trained medical doctors and may not have an adequate amount of experience in medicine.

For situations like these, we propose a virtual medical assistant that guides a caregiver through a procedure by communicating with the caregiver in order to complete the procedure successfully. A successful procedure includes maintaining patient safety and lowering caregiver stress while working efficiently (Manser, 2009).

When guiding human beings during a stressful situation, communication is key. The agent system must choose its words wisely in order to guide the human through each step of the procedure. The human caregiver must be motivated by the agent and given an appropriate amount of instruction. We aim to accomplish this by using situational leadership, a management model in which leader communication depends on follower readiness (Hersey et al., 1988). In this scenario, the follower is the human caregiver who follows the direction from the agent, the leader. Follower readiness is defined by a combination of either low or high ability (or competence) and low or high willingness (or confidence). The leadership communication and behavior is chosen based on the follower readiness profile, described in the next section.

While our agent utilises multi-modal behavior (Collins Jackson et al., 2019; Collins Jackson et al., 2020), verbal interaction is the main modality. In this work, we investigate how situational leadership can

[a] https://orcid.org/0000-0002-4385-5258
[b] https://orcid.org/0000-0003-1717-6048
[c] https://orcid.org/0000-0002-7117-3748

Jackson, A., Glémarec, Y., Bevacqua, E., De Loor, P. and Querrec, R.
Speech Perception and Implementation in a Virtual Medical Assistant.
DOI: 10.5220/0010914500003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 1*, pages 106-117
ISBN: 978-989-758-547-0; ISSN: 2184-433X

Table 1: The follower readiness levels (denoted with *R*) and their corresponding leadership styles.

Follower readiness level	Leadership style
(R1) Low ability and low willingness	→ (Directing) high task and low relationship behavior
(R2) Low to some ability and high willingness	→ (Coaching) high task and high relationship behavior
(R3) Some to high ability and variable willingness	→ (Supporting) low task and high relationship behavior
(R4) High ability and high willingness	→ (Delegating) low task and low relationship behavior

be conveyed through semantics and syntax, how differing semantics and syntax are perceived by human beings, and how these perceptions manifest in terms of caregiver performance during a medical procedure.

The rest of this section describes situational leadership, how communication works in a medical environment, how agent communication has been handled in the past, and how all of these things come together in our work to create personalized speech for a virtual medical assistant.

1.1 Situational Leadership

The follower's general readiness to complete the procedure is defined by readiness level which is composed of ability (task competence) and willingness (confidence and interest in completing a task). Readiness level is determined by (1) a number of behaviors that are exhibited during the procedure that dictate their ability and willingness and (2) their performance in terms of ability and willingness in the past (Hersey et al., 1988; Bosse et al., 2017; Collins Jackson et al., 2021a). Leadership style and communication then consists of either low or high task behavior (telling what, when, how, and where to do a task) and low or high relationship behavior (motivating and offering support) (Hersey et al., 1988). Table 1 expresses how follower readiness level relates to leadership style.

For situational leadership to work, the leader must help the followers progress through each readiness level (R1-R4). This means that the communication coming from the leader needs to instill further competence and confidence in the follower in order for them to move onto the next readiness level. Low relationship behavior is designated for followers with low ability and low willingness as they may not be interested in being encouraged. However, the leader performs high relationship behavior for followers with high ability and low to some willingness because their confidence needs to match their high ability level in order for them to progress (Hersey et al., 1988).

1.2 Communication in a Medical Context

Because this work involves a virtual agent assuming the role of an emergency room leader, communication from the agent cannot be unstructured as that can raise safety concerns (Bickmore et al., 2018). However, communication and speech in particular in situational leadership has not been studied before. This research addresses that gap by studying how differences in syntax and vocabulary can express the differences between each leadership style and how those speech differences affect followers' behavior.

Even though speech specifically has not been explored before, situational leadership has been used in supervisor-student relationships (Lerstrom, 2008), in manager-employee relationships (Thompson and Glasø, 2018), and in clinical supervision (Bedford and Gehlert, 2013). The consensus is that situational leadership provides an excellent framework for handling relationships between leaders and followers, although it requires that leaders be fully equipped to deal with a variety of different followers and situations.

Exploring leadership speech in the context of a virtual medical assistant begins with examining human leadership in the emergency room. The leader during a medical emergency is the coordinator or the surgeon, the individual who coordinates procedural tasks (Moher et al., 1992; Forster et al., 2005). A successful medical leader is one who interacts in a respectful and helpful way with the members of the team and also directs the team towards the best outcome for the task (Hjortdahl et al., 2009; Moss et al., 2002). The caregivers must have trust in the leader to make the right decisions and ensure that all processes are completed efficiently and correctly (Kulms and Kopp, 2016).

There are several taxonomies that provide guidelines for leader behavior in the emergency room, such as the Non-Technical Skills for Surgeons (NOTSS) (Yule et al., 2008), the Anesthesiologists Non-Technical Skills (ANTS) (Flin et al., 2010), and the Surgeons' Leadership Inventory (SLI) (Henrickson et al., 2013). Elements of these taxonomies were brought together under situational leadership in order to demonstrate how communication and other non-technical skills in the emergency room can be personalized depending on follower readiness level (Collins Jackson et al., 2020).

1.3 Agent Communication

In regards to agent-specific communication, virtual agents have been introduced in medicine in the form of a hospital companion (Montenegro et al., 2019), which provided companionship, information to patients, as well a diagnostic tool (Philip et al., 2020), in which an agents conversed with a human in order to make a mental health diagnosis. In these systems, agents acted as information sources, and their speech was scripted and/or followed a set string of questions. The agents' communication styles did not vary according to the profile of the humans they conversed with. Other work that involves virtual agents cooperating with humans to accomplish tasks includes games in which an agent and a human must work as a team to win a game of basketball (Lala et al., 2015). In this work, the researchers found that communication was hugely important, even more important than agent competence.

The prior work that is closest to our own is that of agents helping humans work during crises. In a simulation of a disaster situation, a planning agent proposed courses of action to first responders (Ramchurn et al., 2015). The agent uses natural language for simply structured orders to the human responders. Their study revealed several things that make cooperation between agents and humans more effective, particularly when the agents are leading the group: (1) adaptivity, meaning that the agent adapts to the needs of the responders; (2) interaction simplicity, meaning that the agent communicates as simply as possible; and (3) flexible autonomy, meaning that the agent allows the responders to control the situation. If we were to classify the kind of leadership that this agent exhibits, it would be delegating leadership. This agent system also takes into account responders' preferences, similar to how work in situational leadership might take into account follower profiles.

Intelligent tutoring systems take these principles of adaptivity, interaction simplicity, and flexible autonomy into account to guide a human through a process and/or teach them something. In these cases, the agent assumes a role of authority. For these situations to work well, the human needs to trust the agent enough to successfully lead him or her through a series of steps (Kulms and Kopp, 2016). Sometimes, embodied tutors take into account the prior knowledge of the user as well as the actions taken by the user throughout the learning experience (Griol et al., 2019) as well as various information states for the agent (Chetty and White, 2019). When the situations are stressful for the humans, adaptivity, simple interactions, and autonomy can lead to more human trust of the agent and therefore more efficiency and success (Kulms and Kopp, 2016).

An agent's personalized content and conversations have been found to improve user engagement, improve the quality of speech, provide timely feedback during the interaction, provide adaptive training, and allow for self-reflection (Kocaballi et al., 2019). Real-time adaptation allows an agent to display believable and socially-appropriate behavior.

This research is especially important as it paves the way for agents to communicate in a wider variety of virtual environments and contexts and explores how agents and humans can communicate for a specific purpose. In each of these examples, the agent communication was hugely important as the communication must facilitate the end goal, whatever that may be.

1.4 A Virtual Medical Assistant

For a virtual assistant agent that guides a human caregiver through a medical procedure, communication must be adaptive, simple, and allow for autonomy when possible. The caregiver may not be experienced in medicine, they may not have experience with the current procedure, and they may not have the confidence necessary to perform certain tasks. The agent must be able to guide followers of any style successfully, which means that the agent's communication may need to change depending on the individual caregiver. Agent speech must be simple enough to be understood by different kinds of people, and it must be straightforward in nature. Finally, the agent system must allow for caregiver autonomy whenever possible in order to allow caregivers to take control of the situation.

Situational leadership lends itself very well to a system that is adaptive, communicates simply, and allows for autonomy. In situational leadership, the leader's communication style varies according to the follower's readiness level. Communication from the agent to the caregiver is simple and straightforward. When the follower has demonstrated an ability to self-lead (readiness levels R3 and R4), the agent is able to take a step back and allow the caregiver some autonomy.

The virtual agent embodying leadership styles exists within a SAIBA-compliant system (Vilhjálmsson et al., 2007). Our current agent framework involves text-to-speech, without an emphasis on intonation. Therefore it is important that all perceptions be determined from text only and without intonation taken into account, hence the reason that only speech in written form was used (although intonation is the subject of future work). More on our agent framework is discussed in section 4.

Despite various studies on the effects of situational leadership, there is a gap in the state of the art in terms of what low-level leader behavior like speech looks like from different leadership styles. In previous work, we explored whether leadership style could be detected from a corpus of leader speech (Collins Jackson et al., 2021b). This past work informed how we created our evaluation (this is further discussed in section 2).

This research explores and ultimately defines different styles of leadership speech in the emergency room and allows a virtual agent to use and adapt such speech. Therefore the work presented here provides novel contributions to the fields of human behavior, healthcare, and intelligent virtual agents.

Our research questions include:

1. Does readiness level influence the perception of task and/or relationship behavior?
2. Does readiness level influence a follower's ability and/or willingness?
3. Is there any correlation between the perception of a leader's task and/or relationship behavior and the follower's ability and/or willingness?
4. Do various characteristics of a sentence influence a follower's ability and/or willingness?
5. Does a follower's performance with regard to ability and/or willingness improve when the follower is matched with the appropriate leadership style?

In this paper, we first detail the user evaluation we conducted in order to explore the research questions above, we analyze the results of that evaluation, we detail how our findings are implemented in our agent framework, and we discuss the conclusions we draw from this work.

2 USER EVALUATION

Our evaluation consisted of multiple sentences from each leadership style evaluated by participants with regards to four different questions asking about their perception of the leader's task behavior, the leader's relationship behavior, their own ability, and their own willingness. In this section, we thoroughly explain how the sentences included were chosen, the design of the experiment, and how participants were recruited and who they are.

2.1 Sentence Selection

In this work, we do not delve deeply into speech acts, but it is worth briefly mentioning them as they explain what the goal of each of the sentences included in this experiment aim to do. When designing agent speech, speech acts are a way of organizing agent speech so that the speech performs the intended effect on the listener. The intended effect on the listener is also called the communication intention (Vilhjálmsson et al., 2007). This work involves an agent leading a human being through a procedure by giving orders to both novices and experts, and so the communication intentions are limited to motivating the human to perform an action with speech acts *directing* in which orders are given and *reporting* in which information about the procedure is stated (Bunt, 2009).

A single context was chosen so that the sentence content itself did not affect responses. We also wanted to choose something that participants generally would feel capable of doing. Therefore, we envisioned a scenario in which the agent needed to motivate the caregiver to disinfect the patient's abdomen. The base sentence is a detailed imperative sentence: "Take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left."

As mentioned in the introduction, our previous work involved exploring the semantics and syntax of medical leader utterances in each leadership style (Collins Jackson et al., 2021b). In that work, a corpus of medical leaders' utterances were assigned leadership style by four annotators. The results imply that there are certain rules that dictate speech in each leadership style. The results in our previous work suggested that while indicators of task behavior were more universal, indicators of relationship behavior were not and should be tailored to individuals' preferences.

These guidelines formed thirty-three sentences that each aimed to have the receiver of that sentence disinfect the patient's abdomen. Of these sentences, thirteen used the guidelines for directing speech, twelve used the guidelines for coaching, two used the guidelines for supporting, and six used the guidelines for delegating. These sentences are described by various attributes:

- Leadership style: which leadership style's guidelines were used to create the sentence;
- Mood: the sentence's mood (imperative, imperative-let in which an imperative begins with the word "Let" (Collins Jackson et al., 2021b), interrogative, indicative);
- Keywords: any relevant keywords that are

present in the sentence (“can”, “could”, “would”, “please”, “I need”, “I want”, “I’d like”, “we”, “help”, “I see”, and “It looks like”);

- Detail level: the level of detail of the instruction (low, moderate, and high);
- Context given: whether an explanation for why the task must be done is present (yes or no).

Some examples of the sentences we used include the following, with a complete list available in Appendix A:

- I need you to prepare the patient by disinfecting the abdomen (directing);
- Could you prepare the patient by disinfecting the abdomen, please? (coaching);
- Do you need any help in preparing the patient? (supporting);
- The patient needs to be prepared before the procedure begins (delegating).

The complete list of these sentences is available in Appendix A.

2.2 Experiment Design

The experiment was conducted as an online survey using Google forms. Each participant was assigned a random readiness level from the list in Table 1 and asked to imagine that they are on a remote site with another human being who has suddenly fallen ill and a third person who is their boss. The participant and their boss must work together to save the patient. The boss is experienced in medicine and has chosen a medical procedure to perform. Participants were told that the boss was also a human being so as to limit the effect that speech from a virtual agent would have. In future work, we test whether the findings from this study hold up during a medical procedure led by a virtual agent.

The participants filled out some demographic information (age, gender, native language, and English level) before moving onto the questions. Each participant answered four questions for each sentence. The questions were in a random order for each participant, and the sentences were randomized for each question. The four questions and possible responses (on a five-point Likert scale) were:

1. Indicate to what extent (from strongly disagree to strongly agree) you agree that your boss is pushing you to do the job *(What is the participant’s perception of the leader’s relationship behavior?)*;
2. Indicate to what extent (from strongly disagree to strongly agree) you agree that your boss trusts you to do the job *(What is the participant’s perception of the leader’s task behavior?)*;
3. Indicate to what extent (from very incapable to very capable) you believe you are capable of completing this task *(What is the participant’s perception of their own ability?)*;
4. Indicate to what extent (from very uncommitted to very committed) you are committed to completing this task *(What is the participant’s perception of their own willingness?)*

2.3 Participants

Participants were recruited through our institution as well as social media and were each entered into a drawing for five 10-euro prizes if they participated.

Eighty-eight people total responded to the survey between October 13th and 26th, 2021. However, one participant responded to every question and every sentence with the same response, so their answers were removed, leaving us with 87 participants.

Participants ranged from 17 to 63 years old (mean = 32.41, SD = 13.30), 43 of which were women, 44 of which were men, and one of which preferred not to report their gender. Over 55% (48) of the participants spoke English as a native language, with French (24 participants), Arabic (5 participants), German (3 participants), Dutch (2 participants), Spanish (2 participants), Italian (1 participant), Polish (1 participant), and Ukrainian (1 participant) making up the rest. Fifty participants responded that they spoke English as a native language (the discrepancy between the number of participants who selected English as their native language and the number who reported that they speak English at a native level may be explained by participants who had sufficient English exposure and therefore determined their English level to be native as well). Regarding the English level of the rest of the participants, 14 self-reported that they were fluent, 20 reported that they were high-conversational, and 3 reported that they were low-conversational. As mentioned, readiness levels were randomly assigned to participants: 22 were assigned to R1, 20 to R2, 25 to R3, and 20 to R4.

3 ANALYSIS

In this section, we present the findings from the user evaluation with respect to the research questions in section 1.4.

Note that we chose not to standardize participant responses. Only four participants’ responses ranged from neutral to strongly agree, very capable, or very committed, and twelve participants' responses ranged from somewhat disagree, somewhat capable, or somewhat capable to strongly agree, very capable, or very committed. Given that there were only four participants who did not select any response indicating disagreement, incapability, or lack of commitment, we decided not to standardize the responses of every participant.

3.1 Influence of Readiness Level on Perception of Task and Relationship Behavior

To address research question 1, we isolated the data to include only responses to one question at a time. We fit a linear mixed effects model with *response* as the outcome variable, with a fixed factor of *sentence*, and a random factor of *readiness level*.

One-way ANOVAs revealed that there was not a statistically significant difference in response to Q1 (ANOVA, p-value = 0.31) or Q2 (ANOVA, p-value = 0.11) between readiness levels, indicating that readiness level had no significant effect on participants' perceptions of task and relationship behavior.

Regarding Q2, there was a statistically significant interaction (ANOVA, p-value < 0.001) between readiness level and sentence as well, meaning that the extent to which the participants felt that the boss trusted them to do the task depended on the combination of the sentence and the readiness level, but readiness level alone did not affect the extent to which participants felt that the boss trusted them.

There were statistically significant effects of gender (p-value = 0.01) and native language (p-value < 0.001) on the perception of relationship behavior. There were also statistically significant effects of gender (p-value < 0.001), age (p-value < 0.001), and native language (p-value < 0.001) on the perception of task behavior. These effects are explored further in subsequent sections, but due to the low number of participants in some categories, we cannot adequately explore these effects with our data.

3.2 Influence of Readiness Level on Ability and Willingness

To address research question 2, we performing the same linear mixed-effects model for Q3 and Q4. There is no statistically significant difference in response to Q3 between readiness levels (ANOVA, p-value = 0.06), indicating that readiness level had no significant effect on participants' perception of their own ability levels. However, there was a statistically significant interaction (p-value < 0.001) between readiness level and sentence.

There was also no statistically significant difference in response to Q4 between readiness levels (ANOVA, p-value = 0.891), meaning that readiness level had no significant effect on participants' willingness to complete the task.

There were, however, statistically significant effects of gender (p-value < 0.001), age (p-value < 0.001), native language (p-value < 0.001), and English level (p-value < 0.001) on the perception of relationship behavior.

There were also statistically significant effects of gender (ANOVA, p-value < 0.001), age (ANOVA, p-value < 0.001), native language (ANOVA, p-value < 0.001), and English level (ANOVA, p-value = 0.01) on the perception of task behavior. Again, these effects are explored further in subsequent sections to find out how participants of different demographics respond differently.

3.3 Correlation between Perception of Task Behavior and Ability

To address research question 3, we evaluate whether the responses to certain questions are correlated. A Kendall correlation (Akoglu, 2018) revealed that there is a moderate positive correlation between the responses to Q2 and Q3 (Kendall τ-b = 0.21, p-value < 0.001). We can interpret this to mean that a participant's perception of task behavior is moderately correlated with their perception of their own ability to do the task. It is possible that perception of the leader's task behavior can influence someone's perception of their own ability.

Among participants who were assigned readiness level R1, the correlation is slightly higher (Kendall τ-b = 0.21, p-value < 0.001), the correlation for R2 participants is higher again (Kendall τ-b = 0.31, p-value < 0.001), the correlation for R3 participants was moderate (Kendall τ-b = 0.25, p-value < 0.001), and the correlation for R4 participants was very weak (Kendall τ-b = 0.07, p-value = 0.04). This indicates that task behavior only affects followers' ability when they are in readiness levels R1, R2, or R3.

3.4 Correlation between Perception of Relationship Behavior and Willingness

Again addressing research question 3, we investigate the relationship between the responses to Q1 and Q4 using Kendall's correlation once again. There was significant evidence to suggest that there was little association between responses to Q1 and Q4 (Kendall τ-b = 0.07, p-value < 0.001). This suggests that participants' perception of relationship behavior is only weakly correlated with their own willingness to do the task.

Among participants who were assigned readiness level R1, the correlation is insignificant although likely nonexistent (Kendall τ-b = 0.05, p-value = 0.10), the correlation for R2 participants is slightly higher (Kendall τ-b = 0.13, p-value < 0.001), the correlation for R3 participants was insignificant again (Kendall τ-b = 0.02, p-value = 0.45), and the correlation for R4 participants was very weak (Kendall τ-b = 0.09, p-value = 0.01). This indicates that relationship behavior only minimally affects followers' willingness.

3.5 Influence of Sentence Characteristics on Ability

To investigate research question 4, we isolated the data to include only responses to Q3 and to each readiness level in order to understand the variables affecting followers' ability in each level. Note that there were singularities between *Context given: Yes* and *Detail level: Moderate*, hence the *NA*s in Table 2.

3.5.1 R1

With the data limited to participants assigned readiness level 1, a simple multiple regression model was fitted with the dependent variable *Response* and the independent variables *leadership style*, *mood*, *keywords*, *detail level*, and *context given*. The majority of variables were found to be insignificant, as shown in Table 2. The best-performing model (adjusted R^2 = 0.15) contained the attributes *context given: no* and *mood: imperative*, with sentences with no context given increasing the base response by 1.18 (p-value < 0.001) and imperative sentences increasing the base response by 1.18 (p-value < 0.001).

When demographic information from participants was added, the model's performance marginally increased (adjusted R^2 = 0.30). Interestingly, native English speakers rated ability lower and German native speakers rated ability higher. Because of the low numbers of participants with certain native languages, the differences of English perception between people from different countries cannot be thoroughly reported on with our data but warrants further exploration.

Gender had the largest impact on participants' perceptions of their own ability. When the non-normal response data was analyzed with a Wilcoxon test, we found that men reported their perceived ability significantly higher than women did (p-value = 0.047), although this may not translate to actual performance, only perception.

3.5.2 R2

Like the R1 data, many of the variables were insignificant (see Table 2). The best-performing multiple linear regression model fitted with data limited to participants assigned readiness level 2 (adjusted R^2 = 0.33) included the attributes *Detail level: moderate*, which increased the participants' perceptions of their ability by 1.02 ($p < 0.001$), and *Detail level: high*, which increased the participants' perceptions of their ability by 1.95 ($p < 0.001$). The model improved again when demographic information was added (adjusted R^2 = 0.40), but no one variable stood out as having a large effect on its own. Again, because of the small number of participants in our study, it is very difficult to point to a certain native language or age range that leads to significant results.

3.5.3 R3 and R4

The multiple linear regression models fitted with data limited to participants assigned readiness levels 3 and 4 did not perform well, and this is likely expected due to the fact that R3 and R4 followers already have high ability. The best-performing models had low values for R^2 (0.06 and 0.02 respectively). That said, ability did increase for R3 participants when the detail level was moderate or high, as shown in Table 2. The model for R4 participants was the only one that was insignificant itself.

3.6 Influence of Sentence Characteristics on Willingness

To continue to investigate research question 4, we perform the same steps we did in section 3.5 but analysing the responses to Q4 instead of Q3. As shown in this section, unfortunately the variables had far less impact on willingness than they did on ability. The best-performing models had adjusted R^2 values of less than 0.04. Demographic information failed to have a significant effect.

That said, there were some interesting findings regardless. For R3 participants, interrogative sentences beginning with *Can* and *Could* marginally increased participants' willingness to complete the task by 0.21 (p-value = 0.049) and 0.24 (p-value = 0.03) respectively. Also, indicative sentences beginning with *It*

Table 2: Linear regression results when the Likert response to Q3 (evaluating participants' perceptions of their own ability) is the dependent variable. Data from each readiness level was examined separately. Singularities in the data led to *NA*s.

	R1		**R2**		**R3**		**R4**	
	Adjusted R^2: 0.14		Adjusted R^2: 0.33		Adjusted R^2: 0.05		Adjusted R^2: 0.01	
	$F(12,713) = 11.08$,		$F(13,646) = 25.71$,		$F(13,811) = 4.25$,		$F(13,646) = 1.64$,	
	p-value < 0.001		p-value < 0.001		p-value < 0.001		p-value 0.07	
	Coef.	*p*-value	Coef.	*p*-value	Coef.	*p*-value	Coef.	*p*-value
(Intercept)	1.60	<**0.001**	1.49	<**0.001**	2.15	<**0.001**	2.22	<**0.001**
Mood: Imperative-Let	0.35	0.33	0.61	0.06	-0.19	0.56	0.03	0.94
Mood: Indicative	-0.06	0.80	-0.11	0.62	-0.12	0.59	-0.10	0.71
Mood: Interrogative	0.35	0.33	0.32	0.34	-0.12	0.73	-0.17	0.69
Detail level: Moderate	0.32	0.07	0.96	<**0.001**	0.55	0.00	0.38	0.06
Detail.level: High	1.27	<**0.001**	1.90	<**0.001**	0.63	<**0.001**	0.087	0.67
Context given: No	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
Context given: Yes	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
Can	-0.57	0.10	-0.16	0.61	0.02	0.95	0.26	0.53
Please	0.10	0.49	0.12	0.35	0.08	0.52	0.07	0.67
Could	-0.60	0.08	-0.16	0.61	0.04	0.90	0.26	0.53
Would	-0.42	0.22	-0.20	0.53	0.14	0.66	0.21	0.61
I need	0.00	1.00	0.22	0.33	0.05	0.81	0.22	0.43
I'd like	-0.02	0.92	0.17	0.45	0.19	0.39	0.25	0.38
It looks like	-0.05	0.88	-0.03	0.92	-0.12	0.67	-0.42	0.23
Help	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
We	0.11	0.49	0.20	0.18	-0.01	0.95	0.01	0.96

looks like had a significant negative effect of -0.72 (*p*-value = 0.01).

3.7 Influence of Matching Leadership Style on Ability

To investigate research question 5, we limited the data to only responses to Q3 and fit a linear mixed-effects model again with a *match* variable which was *yes* if the participant's readiness level and the sentence's leadership style matched. We found that there was a significant difference between followers matched with the correct leadership style and those who were not (ANOVA, *p*-value = 0.01). Using a Wilcoxon test, followers matched with the correct leadership style perceived their ability to be significantly higher than those who were not matched with the correct leadership style (*p*-value = 0.01).

3.8 Influence of Matching Leadership Style on Willingness

Again to investigate research question 5, we performed the same steps that were done in section 3.7 but including only responses to Q4. There was no significant difference in perceived willingness between followers who were matched with the correct leadership style and those who were not (*p*-value = 0.52). This mirrors our results from section 3.6 in which we found that there were few attributes that contributed to followers' willingness.

4 IMPLEMENTATION

Using the results from section 3, we have compiled a list of rules, determined by the results of our evaluation, that should be used when creating agent speech in each leadership. The goal is to increase followers' ability and willingness in each readiness level. Keep in mind that leadership style is matched to readiness level, so the rules listed under *Directing* leadership are tailored to followers with readiness level R1, etc. As discussed in section 3, there are not many conclusions to be drawn in terms of increasing followers' willingness. This is the subject of future work. Our findings along with those are shown in Table 3.

The agent framework is built within Mascaret, a metamodel for an informed intelligent virtual environment in which an embodied virtual human can interact with a user (Querrec et al., 2018). Within Mascaret, procedures such as medical procedures can be formalized and simulated, action by action, in the virtual environment. The procedure actions are specified as states and tools are specified as available resources within the environment. The user can follow the procedure in the virtual environment, and they can interact with virtual humans who provide assistance.

Table 3: A list of guidelines for speech in each leadership style from our evaluation.

	Directing	**Coaching**	**Supporting**	**Delegating**
Task behavior	high	high	high	low
Mood	Imperatives without “let”, Indicatives, Interrogatives	Interrogatives, Indicatives	Indicatives, Interrogatives	Indicatives
Context given	no	no	yes	yes
Detail given	high	moderate-high	moderate-high	low-moderate
Keywords	“We need to, “I want you to”, “Carry on with”, “I need”, “can”, “could”, “would”	“please”, “can”, “could”, “would”	“can”, “could”, “would”, “please”	“I see that”, ~~“It looks like”~~

Virtual humans in Mascaret are embodied conversational agents compliant with the SAIBA framework (Vilhjálmsson et al., 2007). Their high-level communication intentions are translated in multimodal behavioral signals which are transformed in animation.

Since Mascaret knows everything about a procedure, it can provide information about all the actions a user does in the virtual environment during the unfolding procedure. We query Mascaret at each user action in order to detect mistakes or accuracy, and we use such insight to compute the user's level of readiness (Collins Jackson et al., 2022). From the user's level of readiness, we determine the virtual assistant's leadership style (Collins Jackson et al., 2021a).

The virtual assistant must communicate with the user throughout the procedure, for example when a new action needs to be performed or when a mistake has been made. We use its computed leadership style to determine the style of speech using the rules we have established in Table 3.

To create agent speech, an intent planner generates the agent's communicative intentions and then codes them in FML (functional markup language), a behavior planner translates communication intentions (what the agent wants the human follower to do) into verbal signals, and a behavior realizer transforms these signals into animation. Our work on agent speech is therefore centered on the intent planner.

Communication intentions are created from the medical procedure stored inside Mascaret. Because Mascaret holds all information about objects and actions that are inside the procedure and objects within the environment, any number of communication intentions stemming from the procedure can be created.

Follower readiness level is calculated from various behaviors (Collins Jackson et al., 2022). Every action that a follower completes perfectly and any errors that may have occurred are used to calculate the follower's readiness level continuously throughout the procedure in real time. Inside the intent planner, readiness level is updated and stored, and leadership style is calculated from readiness level (Collins Jackson et al., 2021a). Depending on the leadership style, the communication intention is transformed into a communication action. This communication action, the sentence uttered by the agent, should take a number of different forms based on the speech rules we have established in Table 3.

To accomplish agent speech that communicates the same communication intention but differs in style depending on leadership style, we identify a number of structures based on our speech rules' characteristics, and each characteristic is put into a list. Items are randomly pulled from these lists to create a sentence that the agent will speak.

As an example, consider a sentence that is created for the agent using directing leadership. Within the pseudocode below, `actionToDo` represents the next action that must be done within the procedure and `sentence` represents the natural content of the communication action.

Using our findings on leadership speech, directing leadership can take the form of imperatives, interrogatives, and indicatives. Therefore, the list of possible moods for directing leadership are:

```
List<string> LS1Moods = new List<string>
  {"imperative", "interrogative", "indicative"};
```

There is no list for imperative sentences because `actionToDo` begins with an infinitive verb (e.g., “Open”, “Check”, etc.) and therefore contains an imperative by default. The lists of possible beginning structures for interrogative and indicative sentences are:

```
List<string> Interrogatives = new List<string>
  {"Can you", "Could you", "Would you"};

List<string> Indicatives = new List<string>
  {"I need you to", "I’d like you to",
  "I want you to", "We need to"};
```

Actions can be expressed as simply `actionToDo`, which expresses the basic action that needs to be done; `actionToDo.Activity.Description`, which expresses the action in addition to any resources that are necessary, and therefore gives the listener more detail; or `actionToDo.Description`, which provides the action, the resources, and more detail regarding how the action should be completed. These variables contain text that is scripted and associated with each activity and resource within Mascaret. A sentence would be created like this:

```
mood = LS1Moods[rnd.Next(0,
 LS1Moods.Count)]

if(mood == "imperative"):
   sentence = actionToDo.Description);
elseif(mood == "interrogative"):
   sentence = Interrogatives
    [rnd.Next(0, Interrogatives.Count)] +
    actionToDo.Description) + "?";
else:
   sentence = Indicatives
    [rnd.Next(0, Indicatives.Count)] +
    actionToDo.Description) + ".";
```

Some example example sentences that would be created under directing leadership include:

- I need you to place the yellow electrode, which is on the table, on the patient's torso.
- Place the yellow electrode, which is on the table, on the patient's torso.
- Can you place the yellow electrode, which is on the table, on the patient's torso?

Note that unstructured dialogue between the caregiver and the agent is not possible. If the caregiver asks a question, the agent is able to respond, but the agent's speech is limited to only the actions and resources within the procedure. To ensure additional safety, the caregiver can decline the agent's help at any time. When the caregiver feels competent without the agent's assistance, they can decline the agent's guidance during the procedure as a whole. In these situations, the agent will only act as a conduit of communication from the team of medical experts.

5 CONCLUSIONS

In this paper, we analyze the results of an evaluation which argue that style of speech should differ between leadership styles to yield the highest-performing and most-willing followers in each readiness level. While situational leadership has been studied and implemented before, low-level behavior such as verbal communication in situational leadership is something that needed more research. Furthermore, situational leadership has not been implemented in a virtual agent system for use in a virtual agent leading a group of humans. This work confirms that utilising characteristics such as mood, detail level, context given, and keywords can lead to higher ability on the follower's part.

Future work may involve investigating more thoroughly the difference between men and women's ability and reactions to a virtual agent leader's speech as well as the differences between people with different native languages, English levels, and ages. We discovered that there were often significant differences between men and women's perception of their own ability. However, it is unclear whether that translates to an actual difference in ability. A future experiment in which followers of different readiness levels and demographics perform a procedure will allow us to analyze their true ability and willingness and determine how a virtual medical assistant can effectively lead a human follower through a medical procedure.

ACKNOWLEDGEMENTS

This work has been carried out within the French project VR-MARS which is funded by the National Agency for Research (ANR).

REFERENCES

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93.

Araszewski, D., Bolzan, M. B., Montezeli, J. H., and Peres, A. M. (2014). The exercising of leadership in the view of emergency room nurses. *Cogitare Enferm*, 19(1):40–8.

Bedford, C. and Gehlert, K. M. (2013). Situational supervision: Applying situational leadership to clinical supervision. *The Clinical Supervisor*, 32(1):56–69.

Bickmore, T., Asadi, R., Ehyaei, A., Fell, H., Henault, L., Intille, S., Quintiliani, L., Shamekhi, A., Trinh, H., Waite, K., Shanahan, C., and Paasche-Orlow, M. K. (2015). Context-Awareness in a Persistent Hospital Companion Agent. In *Proceedings of the 15th International Conference on Intelligent Virtual Agents*, IVA, pages 332–342, Delft, The Netherlands. ACM.

Bickmore, T. W., Trinh, H., Olafsson, S., O'Leary, T. K., Asadi, R., Rickles, N. M., and Cruz, R. (2018). Patient and consumer safety risks when using conversational assistants for medical information: An observational study of siri, alexa, and google assistant. *J Med Internet Res*, 20(9):e11510.

Bosse, T., Duell, R., Memon, Z. A., Treur, J., and van der Wal, N. (2017). Computational model-based design of leadership support based on situational leadership theory. *SIMULATION: Transactions of The Society for Modeling and Simulation International*, 93(7).

Bunt, H. (2009). The dit++ taxanomy for functional dialogue markup. In *Proceedings of 8th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS, Budapest, Hungary. ACM.

Chetty, G. and White, M. (2019). Embodied conversational agents and interactive virtual humans for training simulators. pages 73–77. The 15th International Conference on Auditory-Visual Speech Processing.

Collins Jackson, A., Bevacqua, E., De Loor, P., and Querrec, R. (2019). Modelling an embodied conversational agent for remote and isolated caregivers on leadership styles. pages 256–259. IVA '19.

Collins Jackson, A., Bevacqua, E., De Loor, P., and Querrec, R. (2021a). A computational interaction model for a virtual medical assistant using situational leadership. WI-IAT, Essendon, VIC, Australia. ACM.

Collins Jackson, A., Bevacqua, E., DeLoor, P., and Querrec, R. (2020). A taxonomy of behavior for a medical coordinator by utilizing leadership styles. pages 532–543. International Conference on Human Behaviour and Scientific Analysis.

Collins Jackson, A., Bevacqua, E., DeLoor, P., and Querrec, R. (2021b). Designing speech with computational linguistics for a virtual medical assistant that uses situational leadership. Human Perspectives on Spoken Human-Machine Interaction '21. FRIAS.

Collins Jackson, A., Gilles, M., Wall, E., Bevacqua, E., De Loor, P., and Querrec, R. (2022). Simulations of a computational model for a virtual medical assistant. ICAART.

Flin, R., Patey, R. E., Glavin, R., and Maran, N. (2010). Anaesthetists' non-technical skills. *BJA: British Journal of Anaesthesia*, 105(1):38–44.

Forster, A. J., Clark, H. D., Menard, A., Depuis, N., Chernish, R., Chandok, N., Khan, A., Letourneau, M., and van Walraven, C. (2005). Effect of a nurse team coordinator on outcomes for hospitalized medicine patients. *The American Journal of Medicine*, 118(10):1148–1153.

Griol, D., de Miguel, A. S., Molina, J. M., and Callejas, Z. (2019). Developing enhanced conversational agents for social virtual worlds. *Neurocomputing*, 354.

Henrickson, P., Flin, R., McKinley, A., and Yule, S. (2013). The surgeons' leadership inventory (sli): a taxonomy and rating system for surgeons' intraoperative leadership skills. *BMJ Simulation and Technology Enhanced Learning*, 205(6):745–751.

Hersey, P., Blanchard, K. H., and Johnson, D. E. (1988). *Management of Organizational Behavior: Leading Human Resources*, chapter Situational Leadership, pages 169–201. Prentice-Hall, 5 edition.

Hjortdahl, M., Ringen, A. H., Naess, A.-C., and Wisborg, T. (2009). Leadership is the essential non-technical skill in the trauma team - results of a qualitative study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 17(48).

Kocaballi, A. B., Berkovsky, S., Quiroz, J. C., and Laranjo, L. (2019). The personalization of conversational agents in health care: Systematic review. *Journal of Medical Internet Research*, 11(21).

Kulms, P. and Kopp, S. (2016). Test. In *Proceedings of the 16th International Conference, Intelligent Virtual Agents*, IVA, pages 75–84, Los Angeles, CA, USA. ACM.

Lala, D., Nitschke, C., and Nishida, T. (2015). User perceptions of communicative and task-competent agents in a virtual basketball game. *ICAART 2015: 7th International Conference on Agents and Artificial Intelligence, Proceedings*, 1:32–43.

Lerstrom, A. (2008). Advising jay: A case study using a situational leadership approach. *NACADA Journal*, 28.

Manser, T. (2009). Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. *Acta Anaesthesiol Scand*, 53(2):143–51.

Moher, D., Weinberg, A. L., Hanlon, R., and Runnalls, K. (1992). Effects of a medical team coordinator on length of hospital stay. *Canadian Medical Association Journal*, 146(4):511–515.

Montenegro, C., Zorrilla, A. L., Olaso, J. M., Santana, R., Justo, R., Lozano, J. A., and Torres, M. I. (2019). A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. *Multimodal Technologies and Interaction*, 3(52).

Moss, J., Xiao, Y., and Zubaidah, S. (2002). The Operating Room Charge Nurse: Coordinator and Communicator. *Journal of the American Medical Informatics Association*, 9(Supplement6):S70–S74.

Philip, P., Dupuy, L., Auriacombe, M., Serre, F., de Sevin, E., Sauteraud, A., and Franchi, J.-A. M. (2020). Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients. *NPJ Digital Medicine*, (1).

Querrec, R., Taoum, J., Nakhal, B., and Bevacqua, E. (2018). Model for verbal interaction between an embodied tutor and a learner in virtual environments. pages 197–202.

Ramchurn, S., Wu, F., Jiang, W., Fischer, J., Reece, S., Roberts, S., Rodden, T., Greenhalgh, C., and Jennings, N. (2015). Human–agent collaboration for disaster response. *AAMAS 2015: Autonomous Agents and Multi-Agent Systems*, 30.

Thompson, G. and Glasø, L. (2018). Situational leadership theory: a test from a leader-follower congruence approach. *Leadership & Organization Development Journal*, 39:574–591.

Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A., Pelachaud, C., Ruttkay, Z., Thórisson, K., Welbergen, H., and Werf, R. (2007). The behavior markup language: Recent developments and challenges. volume 4722, pages 99–111.

Yule, S., Flin, R., Maran, N., Rowley, D., Youngson, G., and Paterson-Brown, S. (2008). Surgeons' non-technical skills in the operating room: Reliability testing of the notss behavior rating system. *World journal of surgery*, 32:548–56.

APPENDIX A

Table 4: The complete list of sentences used in the user evaluation and their leadership styles.

1	I need you to prepare the patient by disinfecting the abdomen.	Directing
2	I need you to take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.	Directing
3	I want you to prepare the patient by disinfecting the abdomen.	Directing
4	I want you to take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.	Directing
5	I'd like you to prepare the patient by disinfecting the abdomen.	Directing
6	I'd like you to take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.	Directing
7	Prepare the patient by disinfecting the abdomen, please.	Directing
8	Take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left, please.	Directing
9	Take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.	Directing
10	We need to prepare the patient by disinfecting the abdomen.	Directing
11	We need to take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.	Directing
12	We will prepare the patient by disinfecting the abdomen.	Directing
13	We will take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left.	Directing
14	Can you prepare the patient by disinfecting the abdomen, please?	Coaching
15	Can you prepare the patient by disinfecting the abdomen?	Coaching
16	Can you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left, please?	Coaching
17	Can you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left?	Coaching
18	Could you prepare the patient by disinfecting the abdomen, please?	Coaching
19	Could you prepare the patient by disinfecting the abdomen?	Coaching
20	Could you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left, please?	Coaching
21	Could you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left?	Coaching
22	Would you prepare the patient by disinfecting the abdomen, please?	Coaching
23	Would you prepare the patient by disinfecting the abdomen?	Coaching
24	Would you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left, please?	Coaching
25	Would you take the antiseptic solution, and disinfect the abdomen by applying it with the cotton balls available to your left?	Coaching
26	Do you need any help in preparing the patient?	Supporting
27	Let me know if you need any help preparing the patient.	Supporting
28	I see that the patient needs to be prepared.	Delegating
29	It looks like the patient needs to be prepared.	Delegating
30	The next step is to prepare the patient.	Delegating
31	The patient needs to be prepared before the procedure begins.	Delegating
32	We are going to begin the procedure soon, and the patient needs to be prepared.	Delegating
33	We are going to prepare the patient.	Delegating