Utterance Behavior of Users While Playing Basketball with a Virtual Teammate

Divesh Lala[1,2], Yuanchao Li[1] and Tatsuya Kawahara[1]
[1] *Graduate School of Informatics, Kyoto University, Kyoto, Japan*
[2] *Japan Society for the Promotion of Science, Tokyo, Japan*
divesh.lala@gmail.com, lyc@sap.i.ist.kyoto-u.ac.jp, kawahara@i.kyoto-u.ac.jp

Keywords: Human-Agent Interaction, Joint Actions, Virtual Basketball, Wizard-of-Oz, Conversation Analysis.

Abstract: Research on human-agent interaction has focused mainly on domains which are conversational in nature, but little work has been done on examining the behavior of interactive agents in domains such as team sports. This paper analyzes utterance behavior in this domain, specifically a virtual basketball game with an agent teammate. The main motivation is to assess the nature of utterances during the course of a game. We use a Wizard-of-Oz system which allows a hidden operator to appropriately respond to user utterances. Utterances are analyzed by annotating and categorizing according to Searle's illocutionary speech acts. We find that there is evidence to support the process of the user beginning with basic utterances needed to play the game, confirming that the agent can understand them, and then moving to more complex utterances. We also find that non-task utterances are used and their proportion increases as the game progresses.

1 INTRODUCTION

Embodied conversational agents (ECAs) have been a major focus for interaction research because face-to-face conversation provides a rich source of phenomena where speech, eye gaze and facial expression can be measured and analyzed. However scenarios such as sports where parties interact by navigating in an open space and use full body movements to engage in collaborative actions cannot be handled by ECAs. Virtual agents which can function in these environments have been identified in previous work (Lala et al., 2014). Aside from sports, other related scenarios include a human-agent team assisting victims across a disaster area or even a human and agent lifting furniture around a house. Autonomous agents which can function in such virtual environments are more feasible than real world robots which do not yet have navigation ability which is on par with humans.

These type of interactions are also of a different nature to conversation. Unlike one-on-one conversation, interactions are relatively infrequent, often repeated, and are used to achieve a shared goal. An example of this is basketball, with the interactions being passing. In terms of utterance and dialog analysis, such interactions have received relatively little focus. The lack of research for this type of agent motivates our work. We wish to create an agent which can recognize speech from the user and interact with them in a natural manner. In order to do this, we require not only a speech corpus but data on the type of utterances used so we can create a dialog model.

The domain of our study is a virtual basketball game with an agent acting as a teammate. The user is able to play the game using only their bodies and without hand-held devices. Our methodology in this work is to conduct a Wizard-of-Oz (WOZ) experiment, annotate and categorize all utterances, then discover patterns *during* gameplay. Unlike conversation we are able to analyze temporal patterns because of the repetitiveness of collaborative actions such as passing.

Understanding temporal behavior of the human is important for agents because they may be better able to infer the human's internal state. For example, at the beginning of an interaction a human may be unfamiliar of how to behave to communicate effectively. As familiarity with the agent increases, the observable human behavior also changes. Such information is crucial for any virtual agent system, not only basketball. Therefore, being able to estimate this sense of familiarity has several implications for the design of agents.

We propose that utterance behavior changes over the course of an interaction. More specifically, we hypothesize that the changes over time are related to the

Lala, D., Li, Y. and Kawahara, T.
Utterance Behavior of Users While Playing Basketball with a Virtual Teammate.
In *Proceedings of the 9th International Conference on Agents and Artificial Intelligence (ICAART 2017) - Volume 1*, pages 28-38
ISBN: 978-989-758-219-6

types of utterances used. At the beginning of the game the human is unsure of the capabilities of the basketball agent in terms of what speech it understands, so will confirm that it can understand commands such as passing. We consider such utterances which contribute directly towards the achieving of the goal of basketball to be **task utterances**. As the interaction progresses, if the agent can prove that it can effectively understand the human, the human starts to experiment with more complex task utterances.

H1 Over the course of an interaction user utterances become more complex, from co-ordinating basic to more complex tasks.

Our second hypothesis concerns utterances which are not task utterances. These can include praising or apologizing to a teammate, which provide evidence that agents are considered as social partners rather than machines. We propose that the proportion of task to non-task utterances decreases over time as the language of users becomes more social towards the agent.

H2 The ratio of task to non-task utterances decreases over the course of an interaction.

We also propose that the utterance behavior of the user has some relationship to the subjective perception of the agent. This utterance behavior is measured in terms of the number of task, non-task, and total utterances from the user, and their perception of the agent is measured through a standard questionnaire, with the dependent measures being intelligence, animacy and likeability. Such a relationship would have implications for human-agent research. If the perception of the agent can be estimated through the utterance behavior of the user, then we have a useful method of user attitude which can be estimated in real-time.

H3 There is a relationship between the frequency of task, non-task and total number of user utterances and the the perception of the agent in terms of intelligence, animacy and likeability.

Our approach to answering these questions is to conduct Wizard-of-Oz experiments using the virtual basketball system described in Section 3. We then describe how we analyze utterances during the experiment by categorizing them according to Searle's illocutionary speech acts. The motivation for this categorization is described in Section 4. From this data we use frequency analyses and questionnaires to address the above research hypotheses.

The contribution of this work is an analysis into the nature of user utterances over time when interacting with a basketball agent. This work can provide guidelines for designing agents which can act appropriately with the user in terms of speech behavior, by knowing what kind of utterances are suitable at particular moments.

2 RELATED WORK

Much research into embodied agents has been related to ECAs. Sophisticated techniques for multi-modal interaction have been able to create ECAs which exist in many specialized and real-world domains such as counseling, job interviews and museum guides (DeVault et al., 2014; Baur et al., 2013; Bickmore et al., 2011) as well as those that partake in more general conversation such as Greta and sensitive artificial listeners (Schroder et al., 2012; Niewiadomski et al., 2009). The purpose of these agents is to engage the user in social interactions primarily through conversation, by using social signals to regulate their behavior.

On the other hand, embodied agents have been developed which engage in a shared virtual task with the user, the earliest being Steve (Rickel and Johnson, 1999; Rickel and Johnson, 2000). These types of agents also communicate with the user through multiple modalities and are often used as training systems. In Steve's case the speech acts were well structured. The focus in our work is on unstructured speech where the user is free to say anything. Joint actions as a basis for communication in teams has been implemented in other work, although this focused on robots or agents which were not humanoid (Li et al., 2015; Bradshaw et al., 2009).

Many studies have analyzed spoken dialog behavior of humans towards virtual agents (Campano et al., 2014; Langlet and Clavel, 2014; Veletsianos, 2012; Robinson et al., 2008; Kopp et al., 2005). These dialogs have been social in nature and any task is largely achieved through conversational means. Task-based systems requiring teamwork arguably contain more command-based language ("Go there", "Pick that up"). Several studies have also investigated social dialog by an ECA in a task-based setting (Veletsianos, 2012; Bickmore and Cassell, 2005; Gulz, 2005), with no clear consensus. It would appear that the value of social dialog in these environments is user-dependent. Furthermore, we could not identify any studies which examine the change in utterance behavior during a single session, which is a main focus of this work.

Real world communication in team sports, including basketball, has also been studied (Poizat et al., 2012; Travassos et al., 2011), but there is limited work on interactive virtual teammates in a sporting domain. Naturally there are basketball video games but com-

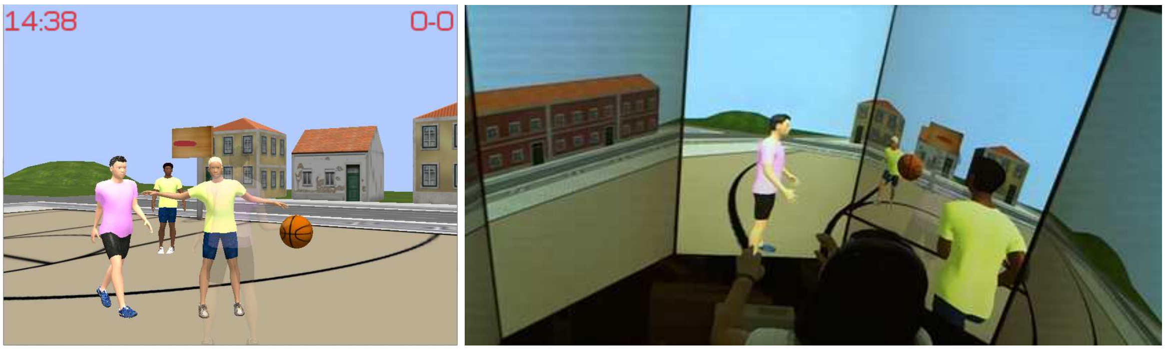

Figure 1: The virtual basketball environment. Screenshot of the game is shown in the left figure while in the right figure the user is shown interacting inside the immersive display environment. A Kinect sensor and pressure pad are used for interaction and navigation purposes.

munication is done through peripherals rather than human body interactions. Furthermore, in a video game the user actually controls all the players, so joint actions are not required. Ideally we would use a robot which could play basketball but this currently does not exist.

3 VIRTUAL BASKETBALL SYSTEM

In this section we describe the virtual basketball environment and the design of the Wizard-of-Oz agent used in the experiment.

3.1 Basketball Environment

Our system is designed so that the user is able to play basketball without the use of keyboard, mouse or hand-held peripherals. This system was also used in previous research to analyze non-verbal signals (Lala et al., 2014) . Our aim is *not* to implement a realistic simulation of actual basketball. This would require overcoming several technical issues which are outside the scope of our work. We concede that the realism of the game can influence the types of utterances used and the results of our study, and this limitation is discussed later in this paper. For now our focus is only on the interactions between a human player and agent teammate, so body movement and speech recognition is required as a means to facilitate natural communication. The actual physics of the game need not be accurately modeled.

The user stands in the middle of an immersive environment, with eight surrounding displays projecting the basketball game. They are represented by a semi-transparent avatar which they can see in a third-person view. The body movements of the user are tracked using a Kinect sensor located in front of them, so gesture recognition of passing, shooting and dribbling can be achieved. To navigate in the environment the user walks in place on top of a foot pressure sensor which recognizes their walking motion and moves their character forward. Due to the limitations of the Kinect sensor, in order to turn in the environment the user does not turn their body but rotates their viewpoint by standing on the extreme edges of the pressure sensor. Although the user must generally be facing towards the Kinect sensor, the immersive displays allow them to perceive the whole of the environment, which is necessary in a dynamic game such as basketball. Screen shots of the game environment are shown in Figure 1. The game is simplified to 2 vs. 2 pickup basketball to stimulate communication. In this version of basketball, each team takes turns at trying to score in one goal only. Opponent agents have the same physical properties as the teammate agent. They will attempt to block the path of the human and find space to shoot goals.

3.2 Wizard-of-Oz Agent Design

Our aim is to eventually create a fully autonomous agent which will recognize human speech. For this reason, we should also use our experiment as a method to collect a number of utterances to create an appropriate speech corpus, which will be in Japanese. We can then use this corpus as a knowledge base for an autonomous agent, and use techniques such as keyword spotting to associate human utterances with behaviors and intentions.

One method of collecting corpus data is to simply observe real basketball matches. However real basketball largely differs from virtual basketball in terms of the richness of communication channels. Real life behaviors make use of facial expression, subtle hand movements and eye gaze which are not recognized in our system. Another approach would be to

Table 1: Categorizations of basketball utterances based on Searle's taxonomy (Searle, 1975). Categories in italics are defined as task utterances.

Illocutionary act	**Utterance category**
Assertive	describing the state of the game
Directive	*calling for a pass*, *ordering (strategy)*, *ordering (shoot)*
Commissive	*throwing a pass*, *statement of intention*
Expressive	*acknowledgment*, apology, celebration, disappointment, encouragement, praise, thanking
Declarative	-
Unclassified	small talk, other

analyze a multi-player basketball game. This also has drawbacks because as humans we can assume many capabilities of each other, including the ability to recognize complex speech. It is likely that most humans will assume their human teammate understands this speech and so use utterances which coordinate human-human activities rather than human-agent play. Research suggests that the type of communication partner (agent or avatar) affects behavior (Fox et al., 2015; Aharoni and Fridlund, 2007).

Due to these issues, we opted to use a Wizard-of-Oz (WOZ) agent. The advantage is that the user assumes that their teammate is artificial while we can provide it with intelligent behavior. The design of the WOZ agent is important because it should not reveal that it is being controlled by a human operator. For this reason the WOZ agent is controlled by keyboard, with triggers for gestures and utterances rather than real-time motion capture and synthesis of a human voice. The utterances are created using OpenJTalk, a Japanese language speech synthesis program (Open JTalk, 2015). This program allows us to create speech for the agent in the form of pre-recorded sound and then playing the sound files during the game at appropriate moments. In total, only 18 sound recordings were used.

The initial utterance categories of the agent were calling for a pass, celebration, disappointment, encouragement and acknowledgment. After the first three experiments we found that we could not encompass a lot of behavior so added new utterance categories in subsequent experiments to help grow our speech corpus. We subsequently added categories of throwing a pass, apologizing and stating an intention to move. The choice of which individual utterances in the same category to use was random. Speech was used to both instigate and respond to the human teammate. For example, the agent could use encouragement if the human was struggling or call for a pass if in free space.

The WOZ operator had knowledge of the goal of the experiment, but their decisions were made to try and simulate those of an average, rational player who aimed to collaborate with their human teammate. Although the game is extremely easy to win using a keyboard, the WOZ operator did not fully realize this capability in order to make the game more balanced.

4 ANNOTATION OF UTTERANCES

In this section we describe the methodology used to annotate and categorize the utterances used by users in the basketball game. Categorization of dialog in human-agent interaction has been addressed in previous research which argued for categorizing dialog based on speech act theory (Traum, 1999; Traum, 2000). However, the majority of this work was in the domain of conversation or conversation as a means of gathering information. The domain of our system is more specific. It is dialog which occurs while a team sport is being played. From our experiments we observed that the type of dialog differed greatly. Utterances tended to be short (two or three words), much like the interactions (one utterance per party), and often repeated at various stages during play. Conversational dialog often involves elaboration, explaining and question-answering as well as facilitation mechanisms such as turn-taking and backchannelling.

It would appear that basketball as a domain is simpler than conversation in terms of the length and type of utterances used. The richness of signaling therefore comes from the context of the game and other modalities. If a player with the ball says “Here!” while turning towards their partner, the partner can infer that this is a signal to receive a pass. Such domains have rarely been examined in real or virtual settings, although research on dialog for online teamwork has been conducted (Taylor, 2012). Therefore we have no domain-specific categorization which we can apply.

Several standardized taxonomies for utterance classification exist and one of the most well-known is the labeling of utterances as illocutionary speech acts as described by Searle (Searle, 1975). Others have also been devised such as DAMSL (Core and Allen, 1997) and DIT++ (Bunt, 2009) which label utterances as dialog acts. These taxonomies address some drawbacks of Searle's categorizations by allowing multiple labels of an utterance and providing a hierarchical structure for categorizations. The dialog acts have been used as the basis for other coding schemes which either refine the tags (Jurafsky et al., 1997) or relate them to specific domains such as meetings (Shriberg et al., 2004).

However a problematic issue with using dialog acts for virtual basketball is that there are many labels to choose from which are applied to human-human conversation rather than basketball-type interaction, as described above. For this reason, we opt to use Searle's speech act categorizations. Although the number of labels is smaller, they better represent the more limited range of utterances used in basketball. Furthermore, annotating and classifying the types of utterances is more clear-cut under the categories defined by Searle as opposed to multi-dimensional or hierarchical labeling. Table 1 displays the categorizations of specific basketball activities under Searle's taxonomy.

We now clarify some of the more ambiguous categories. **Describing the state of the game** is an utterance from the user which explains the current situation but does not make any subjective assessment, such as "*haittenai* ([the ball] didn't go in)". **Ordering (strategy)** is an utterance detailing steps the agent should take in the game, such as moving to a particular location. This excludes passing or shooting commands. **Passing (calling and throwing)** and **ordering (shoot)** were designated as specific categories due to them being the major task behaviors in basketball. **Encouragement** is a general category containing utterances which are used to give the agent support. We include utterances used when the agent is attempting to perform a task or expressing regret for a mistake. Examples include "*ganbare* (do your best)!" and "*ii yo* (it's OK)".

Previously we described task utterances as those being directly related to the achieving of a shared goal. We can therefore also label all commissive and directive speech acts in addition to acknowledgment as task utterances because these are said in order to win the game.

5 EXPERIMENT

We conducted experiments with 15 Japanese speakers who played the basketball game with a Wizard-of-Oz operator. Prior to the game they were shown an instruction video and given a training session to familiarize themselves with the game. During this training session the agent would also take part and engage in a greeting with the participant. This was to ensure that the participants were aware that the agent had the ability to understand speech. We did not provide details as to what speech the participants should use during the experiment. They were free to speak and interact with the agent however they liked. Each game lasted 15 minutes. All speech data and game data (positions of the game objects, players, and their body poses) was recorded so that we could go back and watch the games. Participants were also asked to submit questionnaires which gave subjective evaluations of the perceived intelligence, animacy and likeability of their teammate (Bartneck et al., 2009).

Each recording of the basketball game was observed and all user utterances were transcribed, both lexical and non-lexical. We used the following process to annotate an utterance:

1. If the utterance is not a communicative act toward the agent, ignore it. This removes self-directed speech. This information is maintained for the corpus but is not part of our analysis at this stage.
2. Label the utterance according to the categories in Section 4. This is subjective but when observing the games the appropriate categorization is generally clear, particularly compared to conversation.
3. Note if the categorization is the first of its kind during the basketball game. For example, if the participant says "Thanks" and a thanking utterance has not been used in the game then this constitutes a new category.
4. Note if the utterance is the first of its kind *within the same category*. We consider similar utterances in different categories to be distinct. For example, "Pass" can be used when either calling for a pass or throwing a pass. Variations of the root word constitute the same utterance. In Japanese we consider the utterance "*Pasu shiro*" to be the same as "*Pasu shite*", the common root being "*Pasu*". Although this is not entirely accurate because of nuance, it is satisfactory for this analysis. We also combine repeated utterances into one utterance, defining repetition if it is spoken within 500 milliseconds with no interruption by the agent.

Interpretation of the meanings of utterances was not difficult due to the context of the utterance being apparent in basketball. Nevertheless, the annotations were also checked by a native Japanese speaker and inter-observer reliability was around 95%. The end result of this is a script consisting of time-stamped utterances by both user and agent, and their associated categories. This provides us with the necessary temporal information for our analysis.

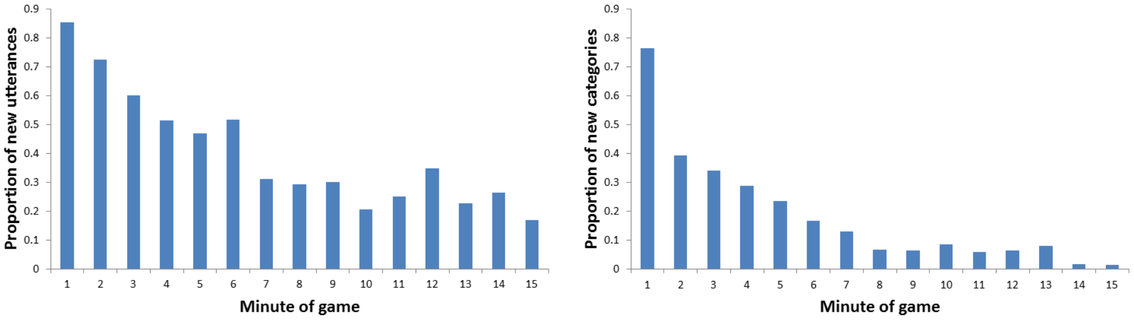

Figure 2: The left figure shows the ratio of new utterances to total utterances divided into 15 1-minute blocks. The right figure displays the ratio of new categories to total utterances.

Table 2: User utterances for all games (abbreviated). Task utterances are in italics.

Utterance category	% total
Call for pass	*18.2*
Praise	13.3
Throw pass	*9.7*
Ordering (strategy)	*9.0*
Acknowledgment	*8.4*
Celebration	6.3
Encouragement	6.3
Apology	5.7
Statement of intent	*4.1*

Table 3: Distribution of utterances under Searle's illocutionary acts.

Illocutionary Act	% total
Assertive	3.5
Directive	31.0
Commissive	13.8
Expressive	47.8
Unclassified	3.9

6 RESULTS

We first provide some general statistics on the utterances. The 15 participants spoke a total of 934 categorized utterances, of which 153 were unique. We identified one outlier, a participant who did not use any utterances during their interaction with the WOZ agent. Table 2 displays the distribution of the utterances for the top ten categories. We can see that utterances are fairly equally spread between task and non-task, although for specific categories, task utterances dominate.

An analysis of the basketball utterances according to Searle's illocutionary speech acts are displayed in Table 3. We see that almost half of all utterances are expressive in nature, while almost a third are directive. This would indicate that communication which expressed emotion or the participant's internal state to the agent was more heavily used than command-based language. Around 15% of utterances were commissive, with the user informing the agent about what they were going to do.

Table 4: Median order of utterance categories.

Utterance category	Order
Call for pass / Throw pass	2
Acknowledgment	4
Celebration / Praise / Thank	5.5
Ordering (shoot) / Disappointment	6
Ordering (strategy)	6.5
Encourage	7
Apology	7.5
Statement of intent	8.5

To analyze temporal behavior we divided the time periods for all participants into 1-minute blocks. The distribution of utterances per block was approximately uniform. There was a mean average of 4.2 utterances per participant per minute. We calculated the proportion of new utterances to total utterances in each block. Results are shown in the left diagram of Figure 2. Until the sixth minute (where there is a peak), the majority of utterances are new. The rate of new utterances then drops after this time and remains fairly steady. We performed the same analysis for the proportion of utterances in new categories, as shown in the right diagram of Figure 2. Similarly, the drop over time is gradual before leveling off from around the eighth minute.

A general overview of the data shows that both task and non-task utterances were used. It would also appear that even after 15 minutes, users would try to sporadically use utterances and dialog with communicative intent which they had not previously used before in the game.

6.1 Task Utterance Complexity

H1 states that users will attempt basic task utterances before complex ones, so we are interested in the order in which new utterance categories are spoken. We analyzed the order of new utterance categories and only considered those which were present in a majority of games, of which there were 12. For example, if "Call for pass" was the first category uttered in a game we recorded its *order* as 1. We took the median of the orders for all games to determine which types of utterance were likely to be spoken before others. The results are shown in Table 4.

What does utterance complexity mean in the context of basketball? Complexity could mean the choice of words used, but as we have stated most utterances were only a few words at most. If we take complexity as the type of action, then the most basic of basketball collaborative actions are to do with passing - asking the agent to receive a pass and signaling that a pass is to be thrown. These can involve both speech and gesture. More complex utterances may be directives which order the agent to perform a particular action or strategy. Results in Table 4 appear to support **H1**. The first utterances from the user are basic passing actions and acknowledgements. Strategic ordering utterances are used later in the game. In terms of Searle's classification, there does not appear to be any definite pattern of ordering.

6.2 Task Utterance Ratio

H2 states that the proportion of task utterances from the user changes during the game. We defined "during the game" as relative to the number of user utterances to account for individual differences in user behavior. To be specific, we divide the total number of utterances of a user into half and compare how the proportion of task to non-task utterances changes over the second half of the interaction. We take a moving average with a large window that can encapsulate a general trend. We first normalized by total utterances, n. For the m^{th} utterance u_m, we calculate the proportion of the previous $m-(n/2)$ utterances which are task utterances $k(u_m)$. This creates a simple moving average for the second half of utterances.

$$k(u_m) = \frac{\sum_{l=m-(n/2)}^{m} 1[u_l = \mathbf{TU}]}{m-(n/2)}, m > n/2$$

with TU indicating whether the utterance is a task utterance.

After normalizing for all users, we then calculated the mean percentage *change* in task utterance proportion during the second half of utterances. Figure 3 displays this trend. It can be seen that the second half of utterances exhibits a decrease in task utterance proportion with approximately 6% less task utterances than the first half. The decrease isn't gradual but fluctuates. This gives some support to our hypothesis **H2**.

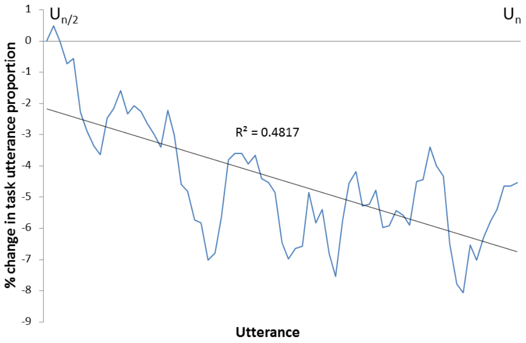

Figure 3: Mean percentage change of task utterance proportion for second half of utterances.

6.3 Perceptions of Agent

We analyzed the results of the Godspeed questionnaire by summing the items for perceived intelligence, animacy and likeability. Cronbach's alpha was above 0.8 for all these measures so we could treat each measure as a single variable. Results in Table 5 show that for all three measures the average score was middling, indicating that users did not generally have strong opinions about the agent. We also found a positive correlation (R-squared value 0.57) between perceived intelligence and likeability.

Table 5: Results of Godspeed questionnaire on perception of agent teammate.

Measure	**Max possible**	**Mean**	**Std. dev.**
Intelligence	25	17.9	3.0
Animacy	15	8.2	2.2
Likeability	25	18.5	2.6

However, we could not find any evidence to support **H3**, which was to assess if there were any correlations between user utterance behavior and their perception of the agent. The number of total utterances, task utterances and non-task utterances had no significant correlation with perceived intelligence, animacy or likeability. From the questionnaires we also found no evidence that users could tell that the agent was actually controlled by a hidden operator. We did not ask the participants about this in the questionnaire itself (to prevent alerting them about the true nature of the experiment), but instead spoke with them casually after the experiment. Several participants remarked on how they were surprised that the agent could understand their utterances.

7 DISCUSSION

Our research assessed utterance behavior during interaction with agents which engage in repeated joint actions with humans. Through a WOZ experiment we were able to produce an agent which could understand the human and react appropriately to a wide range of utterances. Our analysis involved annotating and categorizing utterances then assessing changes over time.

We showed that users tended to begin with utterances which confirmed the agent's understanding of basic passing tasks, before moving on to more complex utterances such as strategic commands (**H1**). We found that subjects used both task and non-task dialog and found some evidence that the proportion of non-task dialog increased during the second half of the interaction (**H2**). There was no evidence that utterance behavior of the user was indicative of their perception of the agent (**H3**). We now discuss limitations of this study and then further discuss these results in a broader context and their implications for future research.

Aside from our hypotheses, we also found that the language of users was varied in terms of the categories of utterances and Searle's taxonomy. This is encouraging because it shows that users did not treat the agent as a simple machine which interacted through commands. In fact, according to Searle's taxonomy, expressive utterances were the most common, with language indicating praise, disappointment, encouragement and apologizing often used. As with real basketball, socially expressive language seems to hold just as much importance in virtual basketball as task-based language. Our hope is this that this type of result can be replicated with an autonomous agent.

7.1 Limitations

There were several limitations in this work. The biggest limitation is that the experiment had a small sample size so our results are only indicative in nature. Although we found evidence of correlations these need to be reproduced to claim any substantial pattern of behavior. In future work, we plan to more robustly test these findings by using more participants and increasing the range of utterances of the agent to accommodate more complex behavior, such as strategic ordering. Furthermore, as we are also aiming to create a Japanese speech corpus, this experiment was performed using Japanese-speaking participants. Cultural or linguistic differences could produce different results in other settings.

One other major limitation is that the game is not exactly the same as real-life basketball. This is not only restricted to physical realism, but also the fidelity of the agent in terms of gaze behavior and body movements. Clearly basketball uses multimodal interaction rather than speech alone. We did not account for these non-verbal features in our analysis, although anecdotally we did observe that users often used non-verbal signals together with speech, particularly when calling for a pass. The agent itself could only utter a very limited set of phrases. This meant that the user could only communicate with it in a limited manner, mainly giving commands and receiving acknowledgments. A more sophisticated agent would need to be able to accommodate small talk behaviors. Additionally, an agent which sounded human-like rather than using a synthesized voice as in this study could have produced better results.

We acknowledge that results of this study could change if a more realistic game was used. However, we also believe that the general hypothesis of communicative behavior shifting from simple to complex meanings would still hold. The difference is the form that this behavior would take, given the ability to smoothly combine speech, gesture and gaze as opposed to reducing signals to speech alone. The challenge is to infer the intended message of the user from a wide range of modalities. Clearly this requires more effort than our study, where only verbal utterances were analyzed. Another challenge for an autonomous agent is to recognize complex multimodal signals, which is more difficult than one modality (in our case, speech recognition).

From the perspective of the user, there was some variation in the ability to play the game smoothly, which may have hindered their motivation to interact. A few users had trouble using the system to pass and shoot which made interaction with the agent troublesome. For these users the focus was on getting the system to work rather than collaborative actions.

7.2 Implications for Agent Design

The long-term goal of our work is to produce an autonomous basketball agent which can interact naturally with the user. However, this does not mean the results cannot be generalized to other domains. Previously we stated that basketball is part of a set of domains which utilize open navigation and full body movement as communicative tools. Another example is helping out victims in a disaster area. We argue that the basic ideas presented in this work still apply to these domains, in that users start by testing basic capabilities of the agent. For a task-based agent these are functions which contribute to the accomplishment of the shared task. Once these have been satisfied, the user is likely to test other capabilities of the agent by engaging in more social language and complex behaviors. We have shown in our experiment that the order of such behaviors can be somewhat estimated. When designing an agent which uses repeated joint actions, we should ensure that we facilitate the user's process of capability testing by creating situations where the agent can prove itself.

This initial experiment can provide a useful baseline for comparison with a fully autonomous agent. We now have a substantially larger corpus from which utterances can be generated, so this should provide a more interesting ground for comparison. Using the corpus we can create an autonomous agent which uses speech recognition. We can then define the utterances the agent uses for specific game actions. From our findings, the agent model should regulate its utterances according to the amount of time spent interacting with the user.

We propose a conceptual agent model based on our findings. During the initial interaction, the agent should show that it can express and understand signals related to simple passing behaviors by actively trying to engage the user in these joint actions. This can be achieved by initiating the joint action through speech and proving to the user it understands this behavior. Several repetitions of these joint actions can be performed. Once this capability has been established, the agent moves to non-task and complex task behaviors using a similar process, gradually building up common ground between itself and the user. With more sophisticated technology, agents and robots which engage in repeated joint actions such as in basketball will become more viable, so we propose analyzing behavior in this environment as a potential research direction.

From our experiment it would appear that user utterance behavior is not correlated with their perception of the agent's perceived intelligence, animacy or likeability. This means that we cannot use real-time utterance analysis in this environment as a means to gauge user enjoyment or satisfaction. It is likely that an analysis of prosodic features of speech such as volume and pitch would produce a correlation with these perception measures, but this requires a more sophisticated recording system than we used for this work. It is also likely that there are non-speech features of the agent which influences user perception.

7.3 Changes in Human Behavior

This work examined behavioral changes over one session of play. In other longitudinal studies with agents, multiple sessions are often used to gauge changes in communicative behavior. We argue that both can be useful, particularly in the case where the same type of interactions occur repeatedly. Although in this work the changes were not drastic even after 15 minutes, we would like to find some underlying state of the user which can be inferred from their behavior. What we could not identify was what causes humans to try new utterances. This information would be extremely useful for agent design because we could use it to speed up the process of the human understanding the capabilities of the agent. The context of our agent makes this crucial because the human must interact with the agent with no prior knowledge of its capabilities.

7.4 Integrating Task and Non-task Dialog

One important result of the experiment is that most participants used dialog which wasn't just directly related to playing basketball. A basketball agent has a particular shared task with a human, as opposed to conversation where the goal might just be to interact socially. The question then arises of how and when to reliably transition from task to social dialog. In the case of basketball, the situations to use both are clearly defined. Task dialog is used during play, while stoppages in play or reactions to an event are precursors to non-task dialog. Social dialog can be considered as a subset of non-task dialog, and is completely unrelated to basketball. For example, during the game an agent may ask if the human plays other sports. We did not find any examples of such utterances in our experiment, but this may need to be considered in the future. After all, many situations such as basketball are essentially tasks which often require social language.

7.5 Comparison with Previous Findings

Our previous work had analyzed body communication during the basketball game. We did not conduct a thorough investigation of body movement in this work, but from casual observations, previous results generally held. Explicit body signals were mainly with the arms and mostly were related to passing interactions. Similarly, both task and non-task communicative signals were used. However in this experiment observable non-task signals such as apologizing and celebrating were done through speech only. One explanation could be that oral communication can express non-task signals much clearer. For example, apologizing without speaking may be unintuitive to humans without detailed recognition of facial expressions. Examination of passing also showed similarities in terms of the initiator and role of the interaction. Participants tended to use speech the most when calling for a pass from the agent, while were less likely to use speech in the opposite situation. In any case, the combination of speech and gesture should be more thoroughly addressed in future work.

8 CONCLUSION

In this paper we analyzed human utterance behavior during interaction with an embodied basketball teammate controlled by a Wizard-of-Oz operator. We found evidence that the utterances from humans toward the agent progressed from coordinating basic tasks to more complex tasks. We also found that humans used both task and non-task utterances, with an increase in the proportion of non-task utterances in the latter half of the interaction. There was no correlation between utterance behavior and the perception of the agent. These results suggest that humans first test if the agent can understand basic speech related to the game before experimenting with more complex joint actions. Non-task dialog should also be considered and be used as the user becomes familiar with the agent. Since we have gathered many utterances for a speech corpus our next step is to create a fully autonomous basketball agent.

REFERENCES

Aharoni, E. and Fridlund, A. J. (2007). Social reactions toward people vs. computers: How mere lables shape interactions. *Computers in human behavior*, 23(5):2175–2189.

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81.

Baur, T., Damian, I., Gebhard, P., Porayska-Pomsta, K., and Andre, E. (2013). A job interview simulation: Social cue-based interaction with a virtual character. In *2013 International Conference on Social Computing (SocialCom)*, pages 220–227.

Bickmore, T. and Cassell, J. (2005). Social dialongue with embodied conversational agents. In *Advances in natural multimodal dialogue systems*, pages 23–54. Springer.

Bickmore, T., Pfeifer, L., and Schulman, D. (2011). Relational agents improve engagement and learning in science museum visitors. In Vilhjálmsson, H., Kopp, S., Marsella, S., and Thórisson, K., editors, *Intelligent Virtual Agents*, volume 6895 of *Lecture Notes in Computer Science*, pages 55–67. Springer Berlin Heidelberg.

Bradshaw, J. M., Feltovich, P., Johnson, M., Breedy, M., Bunch, L., Eskridge, T., Jung, H., Lott, J., Uszok, A., and van Diggelen, J. (2009). From tools to teammates: Joint activity in human-agent-robot teams. In *Human Centered Design*, pages 935–944. Springer.

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Campano, S., Durand, J., and Clavel, C. (2014). Comparative analysis of verbal alignment in human-human and human-agent interactions. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*.

Core, M. G. and Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., et al. (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems.

Fox, J., Ahn, S. J. G., Janssen, J. H., Yeykelis, L., Segovia, K. Y., and Bailenson, J. N. (2015). Avatars versus agents: A meta-analysis quantifying the effect of agency on social influence. *Human-Computer Interaction*, 30(5):401–432.

Gulz, A. (2005). Social enrichment by virtual characters differential benefits. *Journal of Computer Assisted Learning*, 21(6):405–418.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*, pages 97–102.

Kopp, S., Gesellensetter, L., Krämer, N. C., and Wachsmuth, I. (2005). A conversational agent as museum guide–design and evaluation of a real-world application. In Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., and Rist, T., editors, *Intelligent Virtual Agents*, pages 329–343. Springer.

Lala, D., Nishida, T., and Mohammad, Y. (2014). A joint activity theory analysis of body interactions in multiplayer virtual basketball. In *Proceedings of the 28th International BCS Human Computer Interaction Conference on HCI 2014 - Sand, Sea and Sky - Holiday HCI*, BCS-HCI '14, pages 62–71. BCS.

Langlet, C. and Clavel, C. (2014). Modelling users attitudinal reactions to the agent utterances: focus on the verbal content. In *5th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals (ES3 2014), Reykjavik, Iceland.*

Li, S., Sun, W., and Miller, T. (2015). Communication in human-agent teams for tasks with joint action. In *COIN 2015: The XIX International Workshop on Co-*

ordination, Organizations, Institutions and Norms in Multiagent Systems, pages 111–126.

Niewiadomski, R., Bevacqua, E., Mancini, M., and Pelachaud, C. (2009). Greta: An interactive expressive ECA system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 1399–1400. International Foundation for Autonomous Agents and Multiagent Systems.

Open JTalk (2015). Open JTalk - HMM-based Text-to-Speech System. http://open-jtalk.sp.nitech.ac.jp/.

Poizat, G., Bourbousson, J., Saury, J., and Sève, C. (2012). Understanding team coordination in doubles table tennis: Joint analysis of first- and third-person data. *Psychology of Sport and Exercise*, 13(5):630 – 639. A Sport Psychology Perspective on Olympians and the Olympic Games.

Rickel, J. and Johnson, W. L. (1999). Virtual humans for team training in virtual reality. In *Proceedings of the Ninth International Conference on Artificial Intelligence in Education*.

Rickel, J. and Johnson, W. L. (2000). Task-oriented collaboration with embodied agents in virtual worlds. *Embodied conversational agents*, pages 95–122.

Robinson, S., Traum, D., Ittycheriah, M., and Henderer, J. (2008). What would you ask a Conversational Agent? Observations of Human-Agent Dialogues in a Museum Setting. In *International Conference on Language Resources and Evaluation (LREC)*.

Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., de Sevin, E., Valstar, M., and Wollmer, M. (2012). Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2):165–183.

Searle, J. R. (1975). A taxonomy of illocutionary acts. In Gunderson, K., editor, *Language, Mind and Knowledge*, pages 344–369. University of Minnesota Press.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. Technical report, DTIC Document.

Taylor, N. (2012). a Silent Team is a Dead Team: Communicative norms in competitive FPS play. In Voorhees, G. A., Call, J., and Whitlock, K., editors, *Guns, Grenades, and Grunts: First-person Shooter Games*, pages 251–275. Bloomsbury Publishing USA.

Traum, D. R. (1999). Speech acts for dialogue agents. In *Foundations of rational agency*, pages 169–201. Springer.

Traum, D. R. (2000). 20 questions on dialogue act taxonomies. *Journal of Semantics*, 17(1):7–30.

Travassos, B., Araújo, D., Vilar, L., and McGarry, T. (2011). Interpersonal coordination and ball dynamics in futsal (indoor football). *Human Movement Science*, 30(6):1245–1259.

Veletsianos, G. (2012). How do learners respond to pedagogical agents that deliver social-oriented non-task messages? Impact on student learning, perceptions, and experiences. *Computers in Human Behavior*, 28(1):275–283.