Taking Turns in Flying with a Virtual Wingman

Pim Nauts, Willem van Doesburg, Emiel Krahmer, and Anita Cremers

TiCC, Tilburg Centre for Cognition & Communication
Tilburg University, The Netherlands
P.O. Box 90153, 5000 LE Tilburg
{P.A.H.Nauts,E.J.Krahmer}@uvt.nl
PCS, Perceptual and Cognitive Systems,
TNO, The Netherlands
P.O. Box 23, 3769 ZG Soesterberg
{willem.vandoesburg,anita.cremers}@tno.nl

Abstract. In this study we investigate miscommunications in interactions between human pilots and a virtual wingman, represented by our virtual agent Ashley. We made an inventory of the type of problems that occur in such interactions using recordings of Ashley in flight briefings with pilots and designed a perception experiment to find evidence of human pilots providing cues on the occurrence of miscommunications. In this experiment, stimuli taken from the recordings are rated by naive participants on successfulness. Results show the largest part of miscommunications concern floor management. Participants are able to correctly assess the success of interactions, thus indicating cues for such judgment are present, though successful interactions are better recognized. Moreover, we see stimulus modality (audio, visual or combined) does not influence the ability of participants to judge the success of the interactions. From these results, we present recommendations for further developing virtual wingmen.

Keywords: Human-machine interaction, turn-taking, floor management, training, simulation, embodied conversational agents, virtual humans.

1 Introduction

In military aviation there is a tendency towards replacing manned aircraft by Unmanned Aerial Vehicles (UAVs) in hybrid teams (a human lead pilot assisted by a UAV). In the future, human wingmen[1] are foreseen to be replaced by cognitive systems (virtual wingmen) that function as autonomous members in flight operations. Since fighter pilots are highly-skilled, highly-trained professionals, replacing a human wingman by a cognitive system puts high demand on the system's ability to resemble the replaced human in both appearance and behavior.

Given that one of the most important aspects of flying missions in teams is reciprocal trust (the basis of which is a clear understanding among team members) the specific challenge of developing a virtual wingman is to provide the wingman with an

[1] A wingman is the subordinate of the lead in a two-man (*two-ship)* formation.

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2011, LNCS 6762, pp. 575–584, 2011.

understanding of how to interact with its human team members, how to avoid potential misunderstandings and how to solve them when they occur. We argue that, to achieve this, it is important to first look closely at how humans detect and repair miscommunications in this particular setting. This study adds to existing literature in problem detection and -repair by investigating the interaction between a highly skilled operator, a pilot, and a virtual interlocutor, his virtual wingman.

1.1 Embodied Cognition

As a pilot, you want to be assured that whatever happens your partner, either your wingman or your lead, will be up to the task at hand without any need for directing extra attention to a clear understanding among partners. Replacing a human wingman by an UAV, presented as a virtual wingman, thus asks for a cognitive system that can autonomously interact with the lead and other partners in flying as effectively as possible. Establishing the trust and confidence needed in aviation requires a virtual wingman to provide the suggestion of human-like communication capabilities [1] and to meet expectations of a pilot with regard to its specific role (e.g. domain knowledge, reliability).

The specific challenge then becomes to 1) understand the expectations raised by the suggestion of human-like communication and the suggestion of a wingman, 2) to understand what miscommunications occur and if they are related to either source of expectation, 3) to design dialogue repair strategies that stimulate acceptance and trust of the embodied cognition of the UAV. It is thus an effort of joining advances in computer science, specifically artificial intelligence, and insights from social sciences, capturing the richness and dynamics of human behavior in cognitive systems [2].

The level of humanness that human-machine interfaces show has increased over the years, from simple and static agents to virtual characters that interface between and bridge virtual and real worlds, e.g. for training and simulation in military contexts [3] [4] [5]. Such ‘virtual humans’ are defined as cognitive systems that look like, act like, and interact with humans but exist in virtual environments [2]. In other words, they are a virtual extension of a real-world entity (e.g. an UAV). Despite reported experiences with virtual humans there is no consensus on how to build an embodied cognitive system that can live up to the promise of human-like conversation.

1.2 Taking Turns

Natural language dialogue is a very important aspect of humanness a virtual wingman needs to master, as conversation is so defining of human interaction. However, human dialogue is not without errors and problems [6]. Such problems are best viewed as miscommunications - whatever the communicator was trying to convey is not understood by the addressee. Assuming miscommunications negatively affect partners’ satisfaction and the perceived effectiveness and trust and control so important to military aviation, these errors should be recognized and handled by a virtual wingman. They should resemble human repair mechanisms, specifically in a way that yields the highest satisfaction to the human partner [7] [8]. Such repair mechanisms, the ability to indicate when communication goes awry and react to these indicators, are an important aspect of our conversational abilities [10]. They are rich by nature and use multiple cues to signify the occurrence of miscommunications (e.g. linguistic features, facial display, timing) [6] [9]. This information is derived from cues in nonverbal channels to control the flow of conversation [11]. With conversational partners posing and answering questions or statements, a sequence of turn-taking is present in the conversation and controlled using these cues, e.g. an end-of-utterance is pre-signaled by the speaker, indicating a current turn is about to be finished [12]. Giving up a turn is often signaled by looking at the addressee [2].

As turn-taking is part of how humans engage in conversation, it is an important characteristic to mimic in a virtual agent - it determines how natural a conversation will be perceived by the human partner. However, it is not clear whether the rules involved in the floor management of casual conversation extend to our specific domain. As such, in order to let virtual wingmen respond appropriately we first need to provide them with an understanding of how to manage the floor or detect error conditions [13] and when to direct extra effort towards human partners' understanding.

In search for a method to provide this understanding we can best look at how humans, specifically pilots, tackle these problems - supposedly by recognizing and processing cues the conversational partners elicit. As shown in other studies (e.g. [12] [14] [15]), perception experiments can reveal whether or not it is evident these cues can be derived from the conversation. We thus investigate if information on the flow of conversation is present in the interactions between pilot and virtual wingman.

To this end we set up a perception experiment based on perceptual judgment by participants. The experiment is preceded by an inventory of type of interactions (both miscommunications and successful interactions) present in interactions between a virtual wingman and a human pilot.

2 Data Collection

The data used in the analysis and experiment are taken from an earlier experiment within a pilot performance program, conducted to evaluate the attitude of military pilots towards working with a virtual wingman. The character used in the experiment, Ashley, is a female virtual agent under development at TNO Perceptual and Cognitive Systems (The Netherlands). An exhaustive description of Ashley's design is beyond the scope of this paper, important in that respect is Ashley is a life-like virtual agent in that she can mimic a genuine human in her appearance and behavior and can engage in dialogue on a level sufficient to test the context in interaction with human pilots. In the previous experiment, fighter pilots from the Dutch Air Force were assigned to fly a mission together with Ashley from front to back, i.e. practicing with Ashley on executing a *lost wingman procedure* with a pre-flight briefing, a (simulated) actual flight and post-flight debriefing. Their task was specifically in the role of the lead pilot with Ashley being their young wingman in training certain basic flight maneuvers and safety procedures.

3 Data Analysis

Data from the aforementioned experiment were first analyzed for occurrences of miscommunications within the conversation, dividing the raw material into separate interactions (a set of turns that belong to answering a single question, including backchannel behavior) and evaluating each interaction for effectiveness. All selected videos shared the viewpoint (pilot visible, side-frontal; Ashley out of frame) and concern the debrief. In over two hours of video (02:03:01) with seven different pilots a total of 394 interactions were observed.

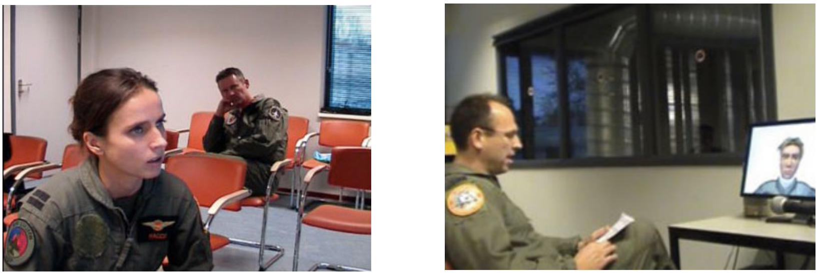

Fig. 1. Left: Example frame from the analyzed videos. People in the background do not engage in the interaction between Ashley and the pilots. Right: The set-up of the briefings. The pilot faces Ashley on the computer screen. A microphone detects speech, speakers are integrated in the room.

3.1 Coding

Each interaction present in the selected recordings was coded and scored from the perspective of the human partner (pilot) as we aim to derive cues from the pilots that might indicate an interaction is failing. Every set of turns, in which a question is posed by one partner and answered by the other, is defined an *interaction.* As strictly coding every set of turns disregards backchannel behavior, backchannel-cues from both partners are included. In most cases, a clear categorization was evident; when in doubt interactions were excluded. See below for an explanation of the coding. Only 'first level' problems are coded, the first occurring problem in an interaction, e.g. if a delayed answer *D* is followed by a nonsensical answer *S,* only the delay is coded. 'Unnatural' in a *Delay* refers to the perception of the coder. As a strict threshold for length (e.g. 3 seconds) disregards context it does not make sense to use an arbitrary measure (a request for simple confirmation requires less time to respond to than does an answer that requires cognitive effort).

3.2 Results

A first indicator of Ashley's performance is the number of positive cases identified. From 394 identified interactions roughly half is successful (***F***, 51.6%). One third of all cases concern delayed responses (***D***, 32.3%), non-responses account for five percent (***N***, 5.4%), interruptions occur in seven percent of cases (***I***, 7.2%) and 3.5% concerns semantics (***S***). We should note these are not balanced results - interactions differ to a large degree on both length and times both partners explicitly contribute within each interaction. Figure 2 above offers a clearer view of the distributions.

Table 1. A technical description of the coding as defined

Label		***Description***
F	*Fluent Interaction*	Turns are appropriately timed and taken, for both question/answers and back-channel cues. Responses are sensical and appropriate
S	*Seemingly fluent.*	Indicates a problem with semantics. Turns are appropriately timed and taken, for both question/answers and back-channel cues. However, responses given are nonsensical and/or out of context.
I	*Interruption*	Wingman takes the turn and interrupts before the utterance is finished or a turn change is indicated.
D	*Delay*	An unnatural delay in the response from the agent after an indicated turn change.
N	*No response*	Virtual wingman does not respond to a question at all, human partner re-takes turn (advances conversation), rephrases or repeats previous statements or verbally indicates it takes too long to respond.

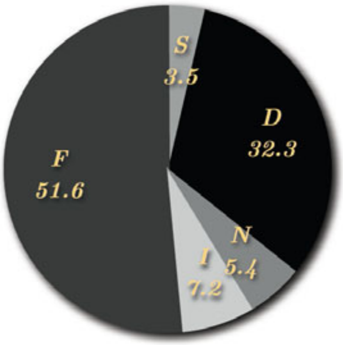

Fig. 2. Qualitative data visualized. Values are percentages. **F** = a successful (fluent) interaction; **S** = Seemingly fluent; **I** = Interruption; **D** = Delay; **N** = Nonsensical.

4 Perception Experiment

Handling error conditions in interacting with a virtual agent requires the agent to understand when these errors occur and how they can be identified. To achieve such understanding, we need to look at how conversational partners indicate problems. The perception experiment was aimed at finding evidence that (1) interactions can be correctly recognized given their success and (2) whether or not certain types of miscommunications yield better recognition over others.

4.1 Materials

For selecting the stimuli from the original materials, only three out of seven pilots in the experiment qualified for a balanced set of stimuli with equally divided successful and unsuccessful stimuli. It is therefore not a statistically representative sample from the data in the source materials. Eight (8x1) successful interactions ***F*** and eight unsuccessful interactions ***N, I, S, D*** (two from each category, 2x4) were selected from the initial data using random sampling, yielding a total of 48 stimuli (3 pilots x 16 stimuli). Using the stimuli, we made three videos (for three conditions) containing sequences of the 48 stimuli accompanied by a written introduction and questionnaires. We manipulated the videos to create three conditions, differing to the degree of richness in communicative cues: an original condition *AV* with both visual and auditory channels and two lean conditions *A* and *V* with auditory- and visual cues removed respectively.

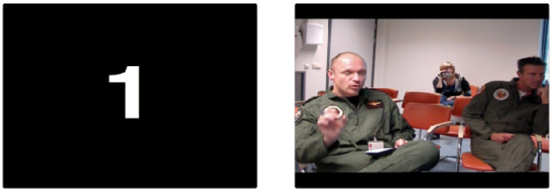

Fig. 3. A screenshot from one of the videos (condition *V*, indicator (1) + a screenshot from the first stimulus on the right).

Sequences in the materials were randomized and balanced by mirroring the sequence to eliminate results in effect of the sequence, yielding a total of six different video compilations (2x3): two different sequences (balance/counterbalance) for each condition (audio, visual, combined).

Questionnaires consisted of a few questions (age, mother tongue, gender) accompanied by a short explanation of the domain and context, a thorough explanation of the task at hand, an instruction and the questions itself. As we wanted to push participants into a decision, a forced choice had to be made: the interaction they were presented with had to be judged in terms of successfulness: each interaction was either a *success* or *no success*. Each question was accompanied by the corresponding number in the video. Participants were seated behind a laptop, and instructed to put on a pair of headphones (for the conditions that contained auditory cues and to block noise for the non-auditory condition) before they randomly received one of six versions of the questionnaires and the corresponding stimuli.

4.2 Participants

A total of 63 students participated in the experiment, 26 males and 37 females aged 18 to 25 (μ = 21.84). Participants were randomly divided over conditions A/V/AV in 21/22/20. No participants or cases were excluded from the analyses.

4.3 Design and Analyses

The experiment was set up using a between-subjects design. The analysis was aimed at the (binary) correctness of the answers participants provided. We first recoded participants' answers into a score for recognition of the stimuli based on our pre-scored attribution in a 3x3x2 design (3 pilots by 3 conditions *A*, *V* & *AV* by 2 categories; *success* & *no success*). In the analysis, *Condition* acted as the independent (between-subjects) variable, measuring its effect on *Recognition* (participants' ability to recognize interactions as successful or not). A second analysis, assessing the effect on the recognition values for *Type* of miscommunication (participants' tendency to recognize some miscommunications more easily than others) did not yield a significant influence of either *Type* on recognition scores. Data were analyzed using ANOVA in a General Linear Model with repeated measures in SPSS Statistics.

4.4 Results

As a first indication, a T-test shows participants are able to correctly recognize interactions above chance levels (t(63) = 8.16, P < .001; M = 29.16, SD = 5.06) (baseline μ_0 = 24, 48 stimuli). The ANOVA shows there are differences in participants' ability to recognize between successful and unsuccessful interactions (F(1, 61) = 68.04, P < .001; η^2 = .527) with success showing significantly higher recognition scores (successful interactions, M = 18.16, SD = 2.81; unsuccessful interactions, M = 13.00, SD = 5.60) (*figure 4, left*). There are clear differences in participants' ability to recognize between pilots (some pilots yield better recognition scores) (F(2, 60) = 22.74, P < .001; η^2= .431). On top of that, we see an interaction effect between *Pilot* and *Category* on participants' ability to recognize (F(2,60) = 17.89, P < .001; η^2= .363), indicating certain types of miscommunications are better observed in certain pilots.

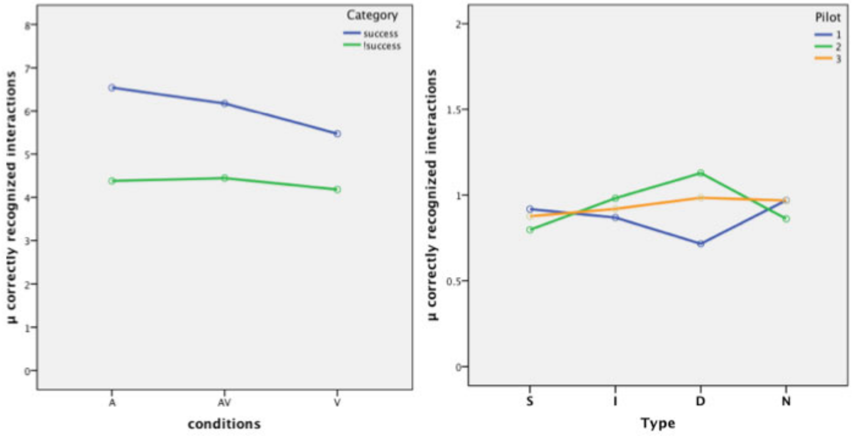

Fig. 4. *Left*: Recognition (Y) scores per category (success | !success) plotted on Condition (X). Y-axis ranges 0 - 8, points represent mean correctly recognized interactions per category per participant for each condition. *Right:* Miscommunications split into Type. (X) plotted for each different pilot. Y-axis ranges 0 - 2, points represent mean correctly recognized interactions per type per pilot. Categories are *Seemingly fluent, Interruptions, Delays, Non-responses*.

There is an interaction effect between *Type* and *Condition*, (F(6,118) = 3.245, P < .001; η^2=.042), indicating differences between conditions on the ability to recognize miscommunications are dependent on *Type* (some miscommunications are more easily observed in certain conditions than others) (*figure 4, right*). No other significant main- or interaction effects were found.

5 Discussion, Conclusions and Recommendations

5.1 Discussion and Conclusions

The descriptives (in section 3.2) show our initial exploration of what characterizes interaction between a virtual wingman and a human pilot in a setting and context that requires a high degree of expertise (i.e. military aviation and operations). As is apparent, Ashley needs to improve her performance by reducing the number of miscommunications - roughly one in two attempts result in miscommunications. However, more than half of these (one-third of the interactions observed) concern delays, relatively easily avoided by improving conversational timing. Moreover, four out of the five types of miscommunications observed relate to floor management. Thus, when Ashley succeeds in appropriate floor management the number of miscommunications can be reduced by almost half (as much as 45%).

In the perception experiment, we observed participants are able to correctly recognize whether a miscommunication occurred or an interaction was successful. This indicates information on such success can be derived from the interaction between a pilot and a virtual wingman based on cues perceivable in the interactions. It supports the idea that cues provided by a human partner can help optimize the interaction between partners in human-machine dialogue [1] [6]. Interestingly, the mean differences indicate participants are better able to recognize successful interactions than they are to recognize miscommunications. This could indicate problems in the interaction are to a certain degree not perceived as such. Surprisingly, the analysis of the observation data shows an insignificant difference in recognition performance between modalities (audiovisual, visual, auditory). This implicates that in designing the module that will take care of floor management, there is no reason other than e.g. system performance to choose for one modality. Furthermore, the second analysis shows there are no differences in recognition between the types of miscommunications - no one miscommunication is more easily observed.

A first limitation of our study is the availability of data and its variance. The number of pilots of which suitable recordings were available is limited and the available data varied to a high degree. With regard to preparing our data, using a non-naive perspective and testing it with naive participants obviously creates a gap - in particular for semantics, where knowledge of the domain is vital to understanding its terminology. For obvious reasons, the coder was not naive to the domain. Moreover, concerning the coding scheme we chose to focus only on first-level problems within the interactions, disregarding the possibility participants might have judged higher-level problems. As a final limitation, in order to succeed in designing natural dialogue for this context, we should incorporate the human partners' attitude towards flying and working with virtual wingmen, not just the bystanders' perspective we investigated.

5.2 Recommendations

Our study can serve as a first exploration of what happens when pilots engage in conversation with virtual wingmen and where for this domain specifically the development of virtual agents should be heading. We suggest improving the ability to perform appropriate floor management in Ashley and other virtual wingmen to succeed in reducing miscommunications. Generally, if virtual wingmen can sense a change in the state of the conversation and understand the minimal threshold involved after which a silence becomes a turn-change, they can construct immediate responses, informing pilots about their state ("uhmmm, let me think"). On top of improving the experience for the human partner, towards more life-like abilities in natural conversation, it would buy wingmen time to process an utterance or construct a response. Importantly, as our study indicates there is a tolerance towards miscommunications in participants' perception, it is advised to take a closer look at how this tolerance extends to military pilots. For other miscommunications beside delays, appropriate repair mechanisms should be designed.

In answer to our research questions, it is (1) indeed evident information on the success of an interaction or turn-taking can be derived from the cues pilots elicit, but to a restricted extent. To improve a wingman's performance, (2) putting effort into capabilities regarding floor management, signaling when turns are given up by the human partner and generating timely responses is advised, as they are the proportionally largest categories of miscommunications that relate to this floor management.

References

[1] Edlund, J., Gustafson, J., Heldnera, M., Hjalmarsson, A.: Towards human-like spoken dialogue systems. Speech Communication 50, 630–645 (2008)

[2] Gratch, J., Rickel, J., Andre, E., Cassell, J., Petajan, E., Badler, N.: Creating interactive virtual humans: some assembly required. IEEE Intelligent Systems 17(4), 54–63 (2002)

[3] Sandercock, J.: Lessons learned for construction of military simulations: A comparison of artificial intelligence to human-controlled agents. DSTOTR-1614, Defence Science and Technology Organisation Systems Sciences Laboratory, Adelaide, South Australia (2004)

[4] Swartout, W., Gratch, J., Hill, A., Hovy, E., Marsella, S., Rickel, J., Traum, D.: Toward virtual humans. AI Magazine 27, 96–108 (2006)

[5] van Doesburg, W., Looije, R., Melder, W., Neerincx, M.: Face to face interaction with an intelligent virtual agent: The effect on learning tactical picture compilation. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 490–491. Springer, Heidelberg (2008)

[6] Martinovsky, B., Traum, D.: The error is the clue: breakdown in human-machin interaction. In: Proceedings of the ISCA Tutorial and Research Workshop Error Handling in Spoken Dialogue Systems. Château d'Oex, Vaud, Switzerland (2003)

[7] Skantze, G.: Exploring human error handling strategies: implications for spoken dialogue systems. In: Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems, pp. 71–76 (2003)

[8] Walker, M.A., Litman, D.J., Kamm, A.A., Abella, A.: Evaluating spoken dialogue agents with paradise: Two case studies. Computer Speech and Language (1998)

[9] Dral, J., Heylen, D.K.J., op den Akker, H.J.A.: Detecting uncertainty in spoken dialogues: an explorative research to the automatic detection of a speakers' uncertainty by using prosodic markers. In: Sentiment analysis: Emotion, Metaphor, Ontology and Terminology, Marrakech, Marocco, May 27, pp. 72–78. ELRA (2008)

[10] Cassell, J.: Embodied conversational interface agents. Communications of the ACM 43(4), 70–78 (2000)

[11] Barkhuysen, P., Krahmer, E., Swerts, M.: The interplay between the auditory and visual modality for end-of-utterance detection. The Journal of the Acoustical Society of America 123(1), 354–365 (2008)

[12] Bulyko, I., Kirchhoff, K., Ostendorf, M., Goldberg, J.: Error-correction detection and response generation in a spoken dialogue system. Speech Communication 45(3), 271–288 (2005)

[13] ter Maat, M., Heylen, D.: Turn management or impression management? In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009, vol. 5773, ch. 51, pp. 467–473. Springer, Berlin (2009)

[14] Swerts, M., Krahmer, E.: Audiovisual prosody and feeling of knowing. Journal of Memory and Language 53(1), 81–94 (2005)