Eindhoven, June 2015

Evaluating multiparty gaze-based turn taking behavior for a robot in situated dialog

by Maurice G.H.H. Spiegels

identity number 0756109

in partial fulfilment of the requirements for the degree of

Master of Science
in Human-Technology Interaction

Supervisors:

Dr. Ir. Raymond H. Cuijpers

Prof. Dr. Wijnand A. IJsselsteijn

ABSTRACT

Advances in artificial intelligence and robotics has increased the interest to develop robots that are designed to interact with people. Nonindustrial or social robots should therefore be capable of human communication traits if they are to be deployed in public services for example. These robots need to be able to make eye contact to create awareness, as well as understand speech to enable natural interaction by verbal dialogue. One of the challenges for conversational robots in verbal interaction, is to avoid interrupting the turn of a human speaker. Conversation analyses reveal that humans have the ability to properly anticipate a turn take moment. This enables smooth floor transitions and results in effective communication. The simplest implementation used by artificial systems to define the end of a turn, is the detection of absence of speech for a predefined fixed duration. Although it is sufficient to facilitate dyadic dialogs between a person and a robot, these systems fall short in multi person dialogs.

The absence of speech is only one of several (aural) cues, and is most often used to determine floor transitions. Tracking the gazing behavior of interlocutors can be considered as an alternative cue, which uses a different modality and can complement the aural cue. However it is not clear how gaze cues can be implemented to improve a dialog system's capability in faster or more reliable turn taking. This study investigated the use of head pose estimation in computer vision to infer people's visual focus of attention, and its value for turn take strategies during small group conversations.

In an experiment, a Nao robot together with participant dyads formed a dialog group. The robot took multiple turns using participant's gaze and absence of speech as cues. With three experimental gaze-based turn take strategies, the perceptual effects of the robot's own responsive gaze as well as variations in turn taking speed on the interaction efficiency were evaluated. To take turn, a conservative strategy required 1 second of silence of the current speaker. This fixed delay allowed for timing analysis of the gaze cue in respect to the absence of speech cue. In contrast, an assertive strategy only needed an instant of silence to take turn. This strategy provided insight to the usefulness of a gaze cue when used in fast turn take strategies. In addition, an adaptive strategy was specified that allowed the robot to take turn without the need for moment of silence. In this strategy a turn took place after a non-fixed delay when a gaze cue was detected. Using a feedback loop the delay was automatically adjusted in an attempt to find a suitable turn take speed.

Turn take strategies that take into account gaze cues are well suited to detect end-of-turns, and provide a way for proper turn taking behavior in multiparty dialog settings. Analyses show that artificial gazing behavior enhances engagement. The timing of the gaze cue in respect to the start of a silence cue is near simultaneous. Substantial spread in timings indicate however that there is no systematic order of appearance of both cues. Gaze cues can therefore increase the robustness of multiparty turn take strategies in combination with the silence cue, but will not necessarily led to faster turn taking. Different turn taking delays significantly influence the way people perceive and evaluate a robot. Despite decreased responsiveness, a more conservative turn take strategy is favored. No suitable turn taking speed was found for the adaptive condition, which led to frequent turn interruptions (cut-ins). In addition, turn taking speed and adequacy had an effect on perceived personality traits of the robot.

Keywords: turn take strategy; speaker transition cues; multiparty verbal interaction; mixed initiative dialog; floor management; behavioral models; responsive gaze; visual focus of attention; speech interface; spoken natural language; multimodal systems; socially assistive robot

Contents

List of figures

List of tables

INTRODUCTION

Human-like artificial systems are expected to play ever more important roles in the domain of public services. Typical employability of these systems is likely to be in environments that provide information or entertainment services, like for example a tour-guide in a museum, a receptionist, or travel agency consultant. Hence, the use of natural language interaction in the near future is likely to dominate over touch interfaces for example. Relevant to the majority of developed interactive robots is the ability to use aural (hearing) and visual (seeing) modalities in order to be aware of actors and changes in their operating environment. Some of the basic skills consist of localizing sound, recognition or production of speech, identification and tracking of human faces, and the ability to estimate people's visual focus of attention. Current technology is powerful enough to replicate several important sensory and perception processes that humans rely on to effectively communicate verbally with one another. Combining these elements enables a spoken dialog system to function reasonably well within a narrow predefined operating environment. For deployment in real-world settings however, current robot communication-behavior models need to be extended to handle situations that involve multiple interlocutors (i.e. group dialog).

People take part in conversations with other individuals every day; sometimes with only one other person, sometimes with multiple persons simultaneously. Conversation in group formations facilitate for example social small-talk, the exchange of ideas in conferences, or propagation of knowledge at councils. Some of the capabilities that are required to partake in natural group conversations is dialog collaboration, tracking others' involvement, and regulating turns. Although technological advances have already resulted in interactive, these systems typically fail in field tests with more than one interlocutor. People as well as artificial systems involved in group communication will therefore not only have to keep track of the information that is being communicated, but also track the dialog flow, turn-changes, and associated cues that enable efficient communication.

Conversational Turn Taking

Conversation analysis studies the verbal and non-verbal aspects of social interactions. A key factor in this research area is the concept of turn taking. The act of taking turns when two or more people have a conversation can be expressed as a mutual symbolic commitment to respect the dialog roles each conversation member chooses to adopt. Goffman (1955) defines turn take signals, which people use to indicate their intention to take over or release the role of speaker. These signals, referred to as turn take cues, are fundamental to facilitate the changes in dialog roles. Important for members of a conversation is to adequately notice and act on these cues. Conversation members must know when a speaker's turn is finished for making the decision to take over the turn to talk, to avoid simultaneous speech or unnecessary silences. Duncan (1972) stresses the significance of speaker turn cues as well. He proposes that only when speaker roles follow from compliance to turn signals, smooth speaker exchanges are possible. Any turn claim attempts by interlocutors while cues are absent, will result in wrongful turn changes often characterized by interruptions and temporal overlap in turns. Speech perceived as interruption during conversation, might however by some people be experienced as mere interjection due to perspective differences on the dialog (Tannen, 2012). Interruptions are therefore only considered erroneous when they are intended to steal the conversation floor, which is called a turn take 'cut-in'. This is regardless whether or not actual speaker overlap occurs. 'Backchannel' responses are a typical phenomenon that often causes speech overlap. These responses are short utterances, for example "uhhums". These are only uttered by a listener to signal that one is still paying attention to what is being said instead of trying to take over the floor (Schegloff, 1982).

Human Turn Take Behavior

Sacks, Schegloff, and Jefferson (1974) defined a model for turn taking that applies to people in conversation. Comprising of only simple statements, the following 3 rules make up the model:

1. The current speaker selects the next speaker who will get the right to speak next.
2. If the current speaker does not use rule 1, the first participant to speak after the current speaker, takes the right to speak (i.e. self-selection).
3. Only if rule 1 and 2 are not used, the current speaker may apply self-selection and resume speaking.

The rules above reflect the mixed-initiative nature of turn taking. The collaborative process is also influenced by multiple factors (Bohus and Horvitz, 2009a). Better turn taking leads to increased dialog effectiveness of which the benefits are often directly perceivable to interlocutors. Most prominently noticeable is the reduced chance of cut-ins, less unnecessary silences due to floor uncertainties, and lesser overall cognitive load that is requested from those involved (Ward, Fuentes, & Vega, 2010). What remains to be defined in the rules above where people select others or themselves to speak, is how this selection is realized during conversation. In addition, there must be a mechanism that makes sure the rules are applied in correct order to prevent speaker ambiguity. Since verbal communication is fundamentally a one-way serial information exchange channel, people need other ways to coordinate turn taking to maintain interaction effectiveness. Several researchers, including those from the human-robot interaction domain, therefor point out to the importance of non-verbal ways of communication such as paralinguistic and behavioral cues to regulate turn taking.

Paralinguistic cues

According to Matsusaka, Fujie, and Kobayashi (2001) people are known to avoid silences if they have the intention to keep the turn. This is done by using meaningless 'filler' utterances (e.g. "sooo…") when one tries to formulate ongoing thoughts. Therefore, absence of speech for a certain duration, from now on referred to as 'silences', most likely indicate the release of the turn. Raux and Eskenazi (2012) complement on this finding by stating that the frequency and duration of silences also correlates well with the mean duration of a person's turn. They also found that at the beginning of a speakers turn, silences that occur tend to be more frequent and have a longer duration compared to silences near the end of a turn. This makes sense since silences near the end of a turn are more prone to be interpreted and used as turn take opportunity by others, and are therefore best avoided or kept to a minimum duration. They also find out that several other prosodic features are of interest to turn taking, such as acoustic energy and pitch of speech which both rise near the end of a speaker's turn. In addition they analyzed the discourse structure and utterance semantics. Words of denial or confirmation like "yes"/"no"/"sure" are found to be reliable end of turn indicators. In particular when these words are preceded by a closed question from a previous speaker.

Gaze cues

Gazing behavior of people is a form of non-verbal body language that in addition to verbal cues plays an important role during conversation. (Kendon, 1967) found that at the beginning of long turns speakers look away; and look back at addressees towards the end of their turn in order to yield the floor. For short turns however, people's gaze tends to be fixed at an addressee during the complete turn without diversions (Cassell, 1998). Incorporating gaze as turn take cue during conversation, can provide the means for designers to increase the speed and/or accuracy of a dialog system. It cannot be assumed however that gazing patterns in human-human interaction also apply to human-robot interaction. This is among other reasons due to the mere fact that current robots cannot be considered fully-fledged conversation partners. Simply a robot's appearance (aesthetics) and spatial factors such as its physical size will have an effect on interactions (Michalowski, Sabanovic, & Simmons, 2006).

Artificial Dialog Systems

A common goal in the development of human-like verbal interaction systems, is to model a robotic system with the ability to understand verbal information, but also to understand the flow of a conversation when multiple people are involved. The dialog flow entails for example who is the speaker, who is addressed by the speaker and is required to listen, and who is expected to be the next one to speak etc. In addition, multiparty settings introduce a problem, which is the difficulty for systems to allocate the source of utterances. Also, the audibility of a system is compromised due to the risk of speech overlap.

Although numerous different system designs are possible, the system defined by Csapo et al. (2012) reflects a very basic but usable solution to accomplish fully fledged functionality on hardware with limited resources. Their dialog-interaction system is divided into three modules. At the heart of the system is the 'conversation manager' that tracks the robot's interactions. It stores various parameters of past-interactions with users and applies a simple set of heuristic rules to ensure richer dialogues. Secondly, the 'Wikipedia manager' (interpreter module) is able to obtain answers from knowledge databases in the form of text (in their case making use of the online free internet encyclopedia Wikipedia). This module makes it possible to accomplish higher levels of artificial intelligence and general applicability. The third 'Nao manager' (behavior module) translates internal actions of the system to perceivable actions in the robot's environment like movements or aural/visual excitations. One of the tasks also belonging to the conversation manager is to regulate turn taking. In collaboration with Csapo and his team, Wilcock (2012) describes in his own paper the use of a finite state machine (flow diagram) to model turn taking. In his design, a finite number of states represent different phases of the robot's behavior for turn taking. The robot's internal system can only be active in one state at a time, and each state can trigger actions of the robot related to that state. As pointed out by Wilcock, state machines have been used successfully for closed-domain dialogs. The type of interaction consists of asking people questions in order to achieve a specific goal such as booking a flight ticket for example in flight reservation systems. The use of states to describe behavior has the advantage of being easier to interpret, and allow for flexibility in the design process.

Modelling Turn Taking

A dialog system has to continuously decide about its own role in conversations, such as being the observant of a human-to-human conversation or being the speaker of a conversation (Matsusaka et al., 2001). Raux and Eskenazi (2009) use a Finite-State Turn Taking Machine (FSTTM) that provides a model with four transitions (actions) between dyadic conversation members that are represented as states. In contrast to the model, the individual actions apply to dyadic as well as multiparty conversations. Used to generally describe the intentions of people during conversation, the same actions are described by Bohus and Horvitz (2010), and are summed in Table 1.

Table 1: Actions in conversation

Keep/Hold	The state of a system when it has the conversational floor, and is either speaking or signaling otherwise to indicate it has the intention to keep the floor.
Wait/Null	The state a system has when it does not have nor claims the floor, and is therefore only observing other parties that are engaged in dialog.
Grab/Take	The system's state when it does not have the floor but attempts to take it. This action (should be) is performed at a well-considered moment in time, based on a turn-take strategy.
Release/Yield	The transition state of the system when it yields the floor, making it available to others after successfully having it for some time.

Even if the system is observing the dialog between other parties, it must make accurate inferences about when it is appropriate to take turn when it 'wants' to say something. Bohus and Horvitz (2011b) emphasize that failures in turn taking will result in perceivable loss of interaction fluidity, which deteriorates the quality of engagement. Coordinating turn take timings is therefore essential to the usefulness and appreciation of a dialog system. The researchers propose that the key to proper turn taking is predicting the finality of a user's turn beforehand, as a way to cope with system latencies that would otherwise limit the responsiveness of a system. Going one step beyond, Ward et al. (2010) point out that predictions of a system's turn take moment should not be made only at certain points in time during the conversation, but should be calculated by the system continuously and ahead of time to realize fast responsiveness. A predictive turn take strategy was for example also used by Raux and Eskenazi (2009), which was based on a decision-theoretic model. The model initiated turn take actions at moments that had the lowest expected costs, which were calculated using probabilistic estimates of the floor's availability. Although results vary, many prediction-based implementations need further development to be successful in multiparty dialog settings, but stay relatively complex or not generalizable/transparent enough (e.g. trained neural network).

Incorporation of multimodal turn take cues might allow for simpler predictive strategies or even non-predictive strategies to be adequate enough for practical use. If a robotic system is able to accurately sense multiple behavioral aspects of human interlocutors, there is no existing framework however that translates this multimodal information to human-like turn taking behavior for a system. This is in contrast to more basic frameworks used to facilitate turn taking, which consist of rules regarding occurrences of silence during conversation. An example is the turn take policy described in the work of Bohus and Horvitz (2011a) for example, in which two rules define the main behavior of their system. The first rule prohibits the system from producing speech when someone else is speaking, to avoid overlap and interruptions, while the second rule enables the system to take the floor whenever during conversation a silence occurs that lasts for a certain duration. Such fixed-threshold strategies require a designer to accept a certain ratio of cut-in risk versus system responsiveness. If the duration of the silence is set short, than cut-ins will likely occur. When requiring a longer duration of the silence, the number of cut-ins will decrease but will also reduce the system responsiveness. This trade-off leads to suboptimal performance, whereas a framework extended with rules concerning multi-model turn take signals may lead to an improvement in performance. Including people's visual focus of attention is likely to be a good extension as it plays an important role in turn take management and the formation of conversational roles (Cassell, 1998). People's visual focus of attention, also known as gaze, is one of the most salient signals in terms of reliability and measurability. More recently Mutlu, Yamaoka, Kanda, Ishiguro, and Hagita (2009) add that, to design more natural and richer interactive robotic behavior, research should focus on communicative rich non-verbal cues, such as people's gaze patterns.

Responsive gaze

For an avatar or robotic system to follow the flow of a conversation using gaze cues, can be considered only half of the work. A robot's own gazing behavior also forms an essential part of the behavioral design. Several studies emphasize that these systems, either physical robotic faces or projections of 3D modelled virtual faces, should also be capable of reciprocal eye contact. A conclusion from research by Sidner, Kidd, Lee, and Lesh (2004) for example, was that people perceived a robot as far less engaging when it did not gaze back at them, as opposed to when it did look at them. An explorative study dedicated to the effects of responsive gaze is done by Yoshikawa, Shinozawa, Ishiguro, Hagita, and Miyamoto (2006). One compelling finding is that participants who interacted with a robot that was capable of maintaining mutual gaze, made participants feel being looked at in contrast to conditions in which the robot did not exhibit responsive gaze.

Extra efforts are often needed to get people familiar with new technologies. When confronted with a speaking artificial system, people appreciate and even assume that these systems will visually respond in intuitive ways (Trafton, Bugajska, Fransen, & Ratwani, 2008). Gazing behavior that is particularly suited for open world interactions in which people move freely and come and go at any given moment is designed by Bennewitz, Faber, Joho, Schreiber, and Behnke (2005a). In their research a robot briefly gazes to non-primary interlocutors like bystanders or overhearers, which was perceived as interest and made them feel involved. At the same time this behavior prevents the robot from staring too long at a speaker, which would be undesirable and unnatural behavior. Bennewitz and her team take a probabilistic approach to maintain the robot's knowledge about its interaction members, such as their position, distance, identity, and level of interest. The gazing behavior comes forth from knowledge uncertainty levels due to the limited field of view of the cameras that the system uses to scan, detect and track people. As a result, the robot will intentionally divert its gaze away from a speaker to 'update' the parameters of other people in its vicinity.

Knowing people's gaze direction and implementing reciprocal gaze, can be useful to interaction systems. New research opportunities might even arise when multiparty artificial conversation systems are further developed. Human accomplices that are used in conversation analysis research for example often have to exhibit certain behavior that requires considerable training time. This holds in particular for studies into the effects of subtle varieties in dialog such as duration of mutual gaze, turn yield ratio, or turn take responsiveness. The capability of a robot to repeatedly show very consistent systematic behavior, could therefore eventually outperform any human accomplice in multiparty conversation studies. Despite the available knowledge, there are no specific details of modelling several cues including gaze to facilitate turn taking in group conversations. It is therefore of interest how such a system could be implemented and what the potential benefits of models are.

The Current Study

The goal is to implement a framework that enables to study a robot's turn take behavior in multiparty verbal interaction. Contributing to the positive experience of a dialog, is the rate of dialog flow which requires minimal delays between speaker transitions and minimal occurrences of speaker overlap (Bohus & Horvitz, 2010). Hence, it is hypothesized that participant evaluations are more positive for turn take models that reduce transition delays and speaker overlaps. In addition, negative evaluations will probably correlate with the frequency and perceived severity of cut-ins caused by the robot when it interrupts a speaker by taking turn. With respect to improper turn-taking, a robot can be considered in the same way as a child that still has to learn when it is acceptable to speak. From intuitive sense it is therefore anticipated that participants are relatively tolerant to turn take errors, with the exception for obvious interruptions that will not be tolerated and probably cause confusion or frustrations.

The usefulness of gaze as a turn taking cue in small group conversations will also be evaluated. Although turn taking is partly regulated by non-verbal signals such as gaze, it is unclear how moments of gaze can be used as turn taking cue. Therefore, the respective timing of the gaze cue with the absence of speech cue will be measured for several turn takes. Participants are likely to favor gazing to each other and look at the robot again only after they have finished mutual discussion, due to the lower status of the robot with respect to human intelligence/conversation capabilities. The gaze cue would in that case be valuable to increase turn take model reliability, but not necessarily lower the required duration of the absence of speech cue.

In addition an assessment will be made on the influence of the robot on conversational dynamics, in relation to its turn take strategies and responsive gaze behavior. Quantitative measures such as the duration of utterances, speaker silences and turn intervals between the two participants, will provide insight into people's speaking behavior that shape the conversation dynamics. The ability for people to adapt to new situations gives reason to believe that more noticeable turn taking strategies may influence the dialog behavior of participants. Since the robot will at all times actively follow and participate in the conversation by turning its head, it is expected that participants will feel personally addressed by inquiries as well as increase ratings of likability and liveliness.

METHOD

Research Design

For this research a Nao robot is used that took part in multiparty dialogs with two participants. The main design principle for the interactive robot was that it actively participated in a conversation. Hence, the robot acted as a travel agent during the experiments, assisting two participants to define a holiday in terms of preferred destination, accommodation, holiday type, budget etc. This task was chosen because of its practical relevance and because it offered multiple opportunities for the robot to take turn.

Inferences of the robot about when to take turn and how, were always based on a combination of two cues; the absence of speech and gaze. The effects and efficiency of different gaze-cue based turn taking strategies were studied using a between-subject experiment design with three conditions. This was because subtle differences between gaze-based turn take strategies might not have yielded enough diversity in data to argue the pros and cons of one strategy over another. Widely dispersed strategies would however eliminate the possibility for any refinement in analyses. In each condition, respectively 'conservative', 'assertive', or 'adaptive', the robot used a different turn take model that mainly influenced the robot's turn take responsiveness (speed) during conversation.

As noted by Hassenzahl, Platz, Burmester, and Lehner (2000) the evaluation of an interaction system is influenced by its hedonic as well as its pragmatic quality, or in other words the non-functional and functional aspects of the robot. By collecting both quantitative and qualitative measures, also known as a 'triangulation design' (Creswell, Plano Clark, Gutmann, & Hanson, 2003), different types of data was collected to evaluate multiple aspects. In short, the three different turn take behaviors enabled comparison of objective measurements such as participant's own turn take behavior (gaze and speech timings), cut-in frequency of the robot, and comparison of subjective factors from questionnaires.

Turn Take Models

Conservative

A 'conservative' condition was defined to gain quantitative data about participant's gaze behavior. It used a long silence cue duration which provided the time needed for accurate measurements as well as minimizing the chance of wrongful turn takes (cut-ins). To set the duration length, inspection of the overall latency versus cut-in tradeoff in the work of Raux and Eskenazi (2012), revealed a useful upper boundary of roughly 1 second that should result in a low cut-in rate < 2%. This means whenever during conversation a speech pause of 1s took place, this was seen as a very reliable cue to take turn. However the robot only took the turn when a gaze cue was additionally present, because in multiparty settings the gaze cue provides information on who should speak next. The relative timing of the gaze and silence cue were not known in advance. Since a moment of gaze towards the robot during the 1 second lasting silence cue would suffice to trigger a turn take, an equally valid but inverse order of cue appearance was also defined by allowing a gaze cue to occur no earlier than 1 second before the start of a silence cue. Hence, a maximum total interval of 2 seconds between the two cues was considered sufficient to successfully demonstrate a multi-cue based turn take strategy. The feasibility of this strategy was also verified during pilot tests.

Assertive

By requiring the presence of a gaze cue, conventional trends in silence cue latencies and associated cut-in rates may not be applicable anymore. Depending on the validity of gaze as turn take cue, and its timing relative to the silence cue, the speed of a turn take strategy could possibly be improved. To test this assumption, the ‘assertive’ strategy formed the counterpart of the conservative condition. It enabled high system responsiveness at the expense of higher cut-in risks. The turn take strategy provided insight to the effectiveness and costs of using faster turn taking by means of primarily the gaze cue. To achieve a turn take strategy that seizes any possibility to take turn, a short lasting silence cue was defined. In several of the earlier mentioned literature, a reoccurring lower boundary for the silence cue duration lies between 400ms to 500ms. These timings are suggested to define a quick system with a just acceptable cut-in risk. For the purpose of the ‘assertive’ turn take strategy, a more extreme boundary was however favored. The decision was therefore made to make both cues equivalent to each other by defining a silence cue duration < 100ms (i.e. instant). This way a turn take was initiated whenever both cues occurred within a narrow range for only a short moment. It is of importance to note that this strategy as well as the conservative strategy were not capable of interrupting on-going speech. There was however a considerable risk of wrongful turn takes whenever speakers were unintentionally silent during their turn, and simultaneously looked at the robot.

Adaptive

For the third condition a turn take strategy was considered necessary that, opposed to the other conditions, was not to gain more qualitative data concerning the two cues. Instead an ‘adaptive’ strategy was defined in an effort to autonomously find an optimal turn take speed. This turn take strategy was therefore allowed to make mistakes in order to correct itself and adapt a variable turn take delay to increase the chance of better turn taking results. This meant that the turn take strategy was allowed to interrupt (cut-in) ongoing speech. Cut-in occurrences together with additional information about speaker silence near the moment of a turn take, were by means of a feedback loop used to adjust the non-fixed turn take delay. Much like people adapt to others in a conversation, it was possible that the flexible turn take behavior would be considered to be better attuned to the dialog behavior of others. On the contrary, various degrees of accuracy (cut-in rates) and responsiveness (turn take speed) were expected, either for the better or the worse. It was therefore interesting to see if the subjective experience of participants in the adaptive condition tended to be similar to the conservative condition or instead tented be more similar to the adaptive condition.

Measures

Quantitative measurements

One way of measuring the human-likeness of a robot's behavior during conversation is by comparing the behavior between human-human interaction with that of the interaction between a robot and a human (Johansson, Skantze, & Gustafson, 2013). The main objective measures of interest were therefore the speech and gaze characteristics of participants during conversation. In order to parameterize speech in group conversations, frequency and timing information concerning gazes towards the robot and detected speech, were made available through the system logs. This was in line with analysis of Bohus and Horvitz (2010) and Heldner, Edlund, Hjalmarsson, and Laskowski (2011). A total of five quantities (variables) were measured; speaker utterance duration, speaker silence duration, turn intervals, turn cut-ins, and turn overlaps. All are illustrated in Figure 1.

The duration of utterance and silences represented the speech properties of individual participants. Turn intervals, turn cut-ins, and turn overlaps provided information concerning turn take transitions between participants. The turn interval indicated the time between the end of the utterance of one speaker and the start of an utterance from a different speaker. A turn cut-in represented the period of time in which the utterance of one speaker got interrupted and finished by the utterance of another speaker. In a turn overlap, the same speaker that started an utterance also finishes it, although a short overlap is involved that occurred with the utterance of another participant. Turn overlaps can therefore be considered failed turn take attempts, backchannel utterances, or even random noise. No distinction can be made afterwards and which makes the interpretation of turn overlaps ambiguous. To get more accurate measures for the frequencies and durations of the other speech properties, turn overlaps were filtered from the raw speech data.

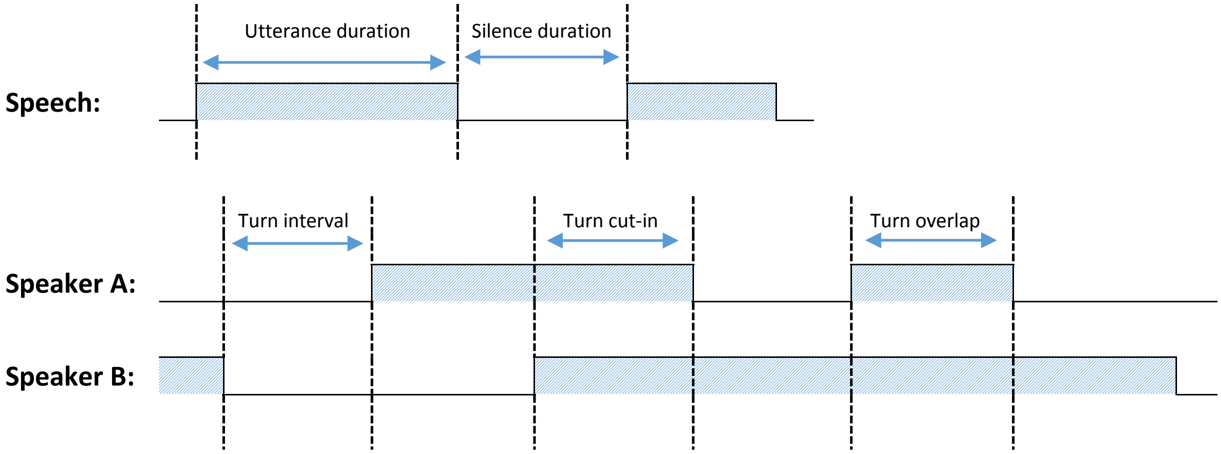
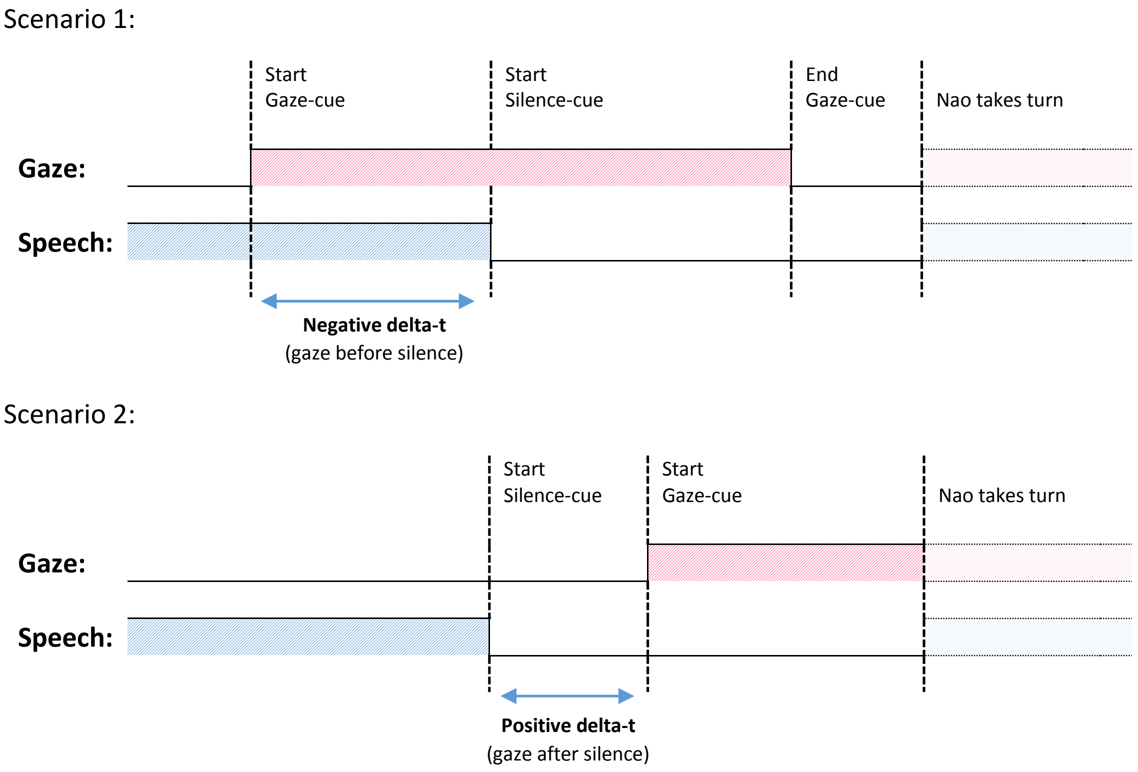


Figure 1: Utterances, silences, and transitions

As important as the mutual relationship between speakers at turn transitions, is the interrelationship between the gaze and speech at turn transitions between a participant and the robot. For the purpose of this study, gaze was considered to be the presence or absence of visual focus-of-attention towards the robot. Any external validation of the head pose estimation was not deemed necessary given adequate measurement results and reliability during pilot tests.

Only applicable for the conservative condition, due to the consistent 1 second silence cue duration, both the relative gaze and silence cue timings could be measured. The variable 'delta-t' indicates whether a gaze cue is received before or after the start of a silence cue. A negative delta-t indicates that the gaze cue precedes the silence cue, while a positive delta-t indicates that the gaze cue occurred after some delay since the silence cue. Both scenarios are depicted in Figure 2.



Scenario 1:

Scenario 2:

Figure 2: Delta-t and Gaze duration

The remaining measures of interest have to do with the adaptive turn take strategy, such as the trend between delay timings and cut-in rates, in combination with the information from custom variables like 'pre-silence' and 'cut-in'.

Qualitative factors of interest

The questionnaire for this study was composed from validated questionnaires that were specifically developed for research in the domain of robotic interaction. Additionally, participants' reactions during interaction with the robot were observed and noted as well. These unstructured notes provided qualitative insight to the experience of participants. See APPENDIX A for the final result, and an overview of the questionnaire's construction and measured factors.

SASSI

The SASSI (Subjective Assessment of Speech System Interfaces) is used to evaluate the usability of the system. It was developed by Hone and Graham (2000) to evaluate the usability of speech-based interfaces for a variety of task-oriented dialog systems, using a seven scaled interval ranging from strongly disagree (−3), to neutral, to strongly agree (+3). Combining the answer score of multiple related items, the questionnaire evaluates 6 different factors which will be briefly described.

1. *Response accuracy*; is measured with 9 items that evaluates how well the system does what the participants expects it to do, and hence relates to correct recognition of the system concerning its input. It should be noted that the level of understanding of the conversations depends not on the system but on the wizard. The system can still however be evaluated for its turn take response accuracy in relation to participants' expectancies and input in the form speech and gaze behavior.
2. *Likeability* (affect); is also measured with 9 items and reflects the opinions and feelings of the participants about the system.
3. *Cognitive demand*; is measured with 5 items. It provides a measurement for the level of effort that participants perceived necessary to interact with the system.
4. *Annoyance*; is measured with 5 items that represent negative associations with the system, and is expected to be inversely related to the likability factor.
5. *Habitability* (transparency); is measured with 4 items and defines the similarity between participant's conceptual understanding of how the system functions and how it actually works.
6. *Speed*; is measured with only 2 items that relate to the system's responsiveness. Given that this factor directly relates to a fundamental difference between experimental conditions in this study, the mere use of two items is expected not to be reliable enough.

Additions to the SASSI questionnaire

Since the type of interaction with the system comprises of verbal dialog, or more precisely natural language interaction, a different questionnaire was involved to resolve any reliability issues. This was the REVU-NL questionnaire (Report on the Enjoyment, Value, and Usability of Natural Language interaction) reported in the paper of Dzikovska, Moore, Steinhauser, and Campbell (2011). They described the development of the survey, and cross validated items with other surveys including the SASSI survey. In order to increase the reliability of SASSI factors due to low item numbers, a total of three appropriate questions from the REVU-NL were added to the SASSI questionnaire. One item, with the highest factor loading on the first component of the REVU-NL, lined up well in their analysis with the SASSI habitability factor, and was therefore added as a 5th item for this factor. In the REVU-NL two items with equally high factor loadings measure system speed. Advisable to use minimally 3 but preferably more items (Raubenheimer, 2004), the two 'speed' questions from the REVU-NL were added to the SASSI, giving it a total of 4 items. Including all changes, a total of 37 questions phrased as statements were used to measure participants' responses. The order of the items were shuffled and inversed randomly, which resulted in 4 versions to eliminate any order effects.

Godspeed

Participant's perception of the robot was measured with the Godspeed questionnaire from Bartneck, Kulić, Croft, and Zoghbi (2009). The user's perception is categorized over 5 factors, of which the first four are measured using 5 items each. Each item is evaluated with a bipolar adjective (e.g. like vs. dislike) on a 5-point scale. Participants are thereby required to indicate which of the two adjectives better fits their experience.

1. *Anthropomorphism*; measures the human-likeness of the robot.
2. *Animacy*; measures how lifelike the robot is.
3. *Likability*; measures the level of appreciation of the robot, which depends on its behavior. Although the same dimension is included in the SASSI questionnaire, the measurement is retained for completeness of the Godspeed questionnaire. Due to the difference in item scales, the two likability scores also provide cross validation of each other's results.
4. *Perceived intelligence*; measures the competence level of the robot perceived by the participants. By using a Wizard-of-Oz design, the intelligence of the robot could be considered compromised. However, turn take behavior of the robot could still affect the perceived intelligence.
5. *Perceived safety*; measures the human's perception of safety during the interaction with the robot. This is the only Godspeed factor that consists of 3 items (instead of 5). This particular measurement is not of special interest for the current research but is retained given that service robots, of which the current robot is an early precursor of, are to be deployed close to people.

Additions to the Godspeed questionnaire

The experiment task for the current study required some social skills of the participants, which may also be expected from the robot. As described by Breazeal (2003), the more socially the behavior of the robot is perceived to be, the more inclined people will be to interact with the robot. This is also generally the case between people that meet. To quantify this influence, a new dimension 'social intelligence' was devised by Mileounis, Cuijpers, and Barakove (2015), consisting of 4 items measuring both the social skills and social competence of the robot. This social intelligence factor was added to the Godspeed questionnaire.

To measure any additional trends that could arise from differences between experiment conditions concerning the robot's turn taking behavior, a final addition to the questionnaire was made by including 8 items with semantic differences. The items follow from terminology used by participants in pilot tests, when they described their interaction experience and thoughts. The items were defined as follows:

1. *Uninterested - Curious*
2. *Impolite - Polite*
3. *Assertive - Conservative*
4. *Distracted - Attentive*
5. *Impatient - Calm*
6. *Dominant - Reserved*
7. *Extrovert - Introvert*
8. *Passive - Active*

Taking into account all modifications, brought the extended Godspeed questionnaire to a total of 35 bipolar adjectives. Again, the order of the items were shuffled, inversed randomly 4 times, and merged with the earlier mentioned 4 versions of the final questionnaire to eliminate any order effects.

Pre-dialogue questionnaire

All participants filled in a short questionnaire providing their demographics (i.e. age, gender, and native language). Game experience was previously found to be significantly related to likability of robots (Mutlu, Forlizzi and Hodgins, 2006). Hence, participants' experience with videogames was measured using a 6-option multiple choice question. In addition, two binary yes/no questions asked participants for any experience with respectively a robot or speech interface system. Participants were also asked for their top 3 expectances regarding the interaction with the robot for the current study.

Post-dialogue questionnaire

After the dialogue, participants first filled in the SASSI and the Godspeed questionnaires. They then wrote down a top 3 of things that struck them most during the experiment task. Combining this information with the expectations of the user with the system asked in the pre-test questionnaire, implicit inferences about the robotic system could be identified. This way of capturing both user expectations and user experiences was based on a method described by Turunen et al. (2009).

Also included in the post-experiment part of the questionnaire were open-ended questions that explicitly asked for any impression, behavior, or perceived personality of the robot. Also an explanation was asked for any affect participants had towards the robot. Finally the participants were asked to write down the presumed goal of the current study, and optionally provide feedback on the experiment. These two questions are a qualitative validation measure for the experiment itself.

Experiment

Setup

To perform the current study, a small humanoid Nao robot manufactured by Aldebaran Robotics was used. For the robot with two participants, equal seating positions around a table was preferred to establish equal interaction angles for all members. Repositioning of participants closer to the robot was however necessary due to computer vision limitations (for technical details, see the MATERIALS chapter). In contrast to placing participants as a dyad in front of the robot, separating them on either side of the robot facilitated more open and mutual conversation. Additionally, this layout required no multi-face detection and enabled the robot to clearly direct its visual attention to either one of the participants. A top view of the setup is illustrated below in Figure 3.

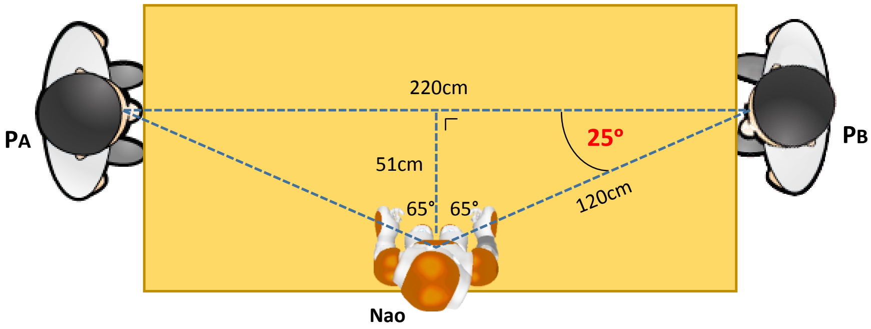


Figure 3: Experiment spatial layout

Task

Inspiration was found in a study that examined multiparty engagement for an avatar in open-world dialogs by (Bohus & Horvitz, 2009a, 2009b, 2010). Their design described playful interaction with an avatar, using a task that is basically a multiple-choice trivia question game. Users respond to questions asked by the system, which in turn provides either verbal feedback about the correctness of the response or provides additional information about the topic of the question. The system was capable of pronouncing utterances to engage bystanders, motivate users to continue playing, and properly switch its attention between participants. Gaming is generally considered to be a fun activity. For the current study, a task was however preferred that did not depend on individuals' knowledge and interest (as many games do), but would still motivate participants to engage in dialog. Also the cognitive load should be low to facilitate easy, more natural conversation behavior. In addition, the task should not require physical objects, in order to prevent distractions as concluded by Johansson et al. (2013) who used a card-based "Desert Survival" team-building game.

This led to the creation of 'The Holiday Planner', in which the robot acted like a travel agent. Given an average duration of 15 minutes, participants had to discuss and find agreement to 11 multiple-choice questions that guided them in defining a hypothetical holiday (e.g. destination, activities, type of accommodation etc.). All topics and associated questions were discussed in pre-defined sequence (phases), to enable a variety of questions that would logically follow from previous given answers. During each phase of the dialog, plenty opportunities were available for the Nao robot to take turn and act as moderator to guide the conversations. The experiment task in combination with the behavioral design of the robot, was aimed to encourages mutual conversation between participants".

Procedure

Participant conversation duos were picked up after spending a short while in the waiting area, and were directed to the experiment lab that looks like an ordinary living room. Once inside, participants were shown the facilities to safely put away any bags or coats before they were appointed to take place at the table on which the Nao robot was positioned. To increase the chance on a successful dialog, participants were properly introduced to become a bit more familiar with each other. In addition an ambient sound system produced non-stop background music at a very gentle decibel level, just enough to eliminate uncomfortable silences while too weak to interfere with the experiment setup. After this brief introduction the participants were asked to read the informed consent form (see APPENDIX G). They were given the opportunity to ask questions about the informed consent to the experimenter. After the signing the form, participants were asked to fill in the first page of the questionnaire, which mainly contained questions about their demographics, robot experience, and expectations. After completion, all table items including the questionnaires were removed from sight to avoid any distractions.

Next the participants got briefed about the experiment in detail. Several dialog topics were provided as example to prepare participants on the content of the conversation, such as holiday countries, activities, period of stay etc. They were told to consult each other for opinions on questions asked by the robot, and to reach mutual agreement for one of the answer options provided by the robot by doing so. In order for the computer vision to work, participants were instructed not to change their current (precalibrated) position at the table. For the same reason large upper body movements were prohibited; small movements of any limbs and head movements to support unrestricted gaze were allowed.

For the success of the experiment, participants were told to immerse themselves in the hypothetical situation that they are going to spend an entire holiday together, and hence the importance of actively starting and participating in the conversation. They were also instructed to avoid very short answers like one-word confirmations as much as possible, and instead use short sentences. In terms of interaction towards the robot, they got informed about its role as moderator and its limitation of not being able to answer any questions they might have. Whenever the robot pronounces an utterance that is not understood, the participants are likewise advised to figure out together what was likely being said instead of asking the robot to repeat itself. In terms of non-verbal feedback from the robot, the participants were told that the 'ear-leds' will turn on at the side(s) of the robot's head that correspond to detected speech of the talking participant on that same side. This provided non-distractive feedback that let the participants know the system was still active and 'listening'.

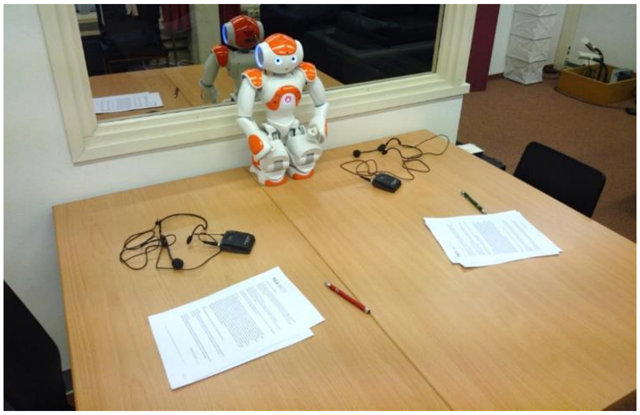


Figure 4: Experiment conversation table

After asking each participant for confirmation that the experiment procedure was clear, the headset microphones were carefully positioned on the participants' heads. Using the built-in decibel gauge, it was possible to adjust the positioning of the microphone. This was needed to eliminate heavy breathing noise that was sometimes picked up by the microphone. After the experimenter took place in the control room, situated behind a one-directional mirror that allowed for observation, the software program was executed that controlled the robot. During the startup phase, calibrations were performed for both the audio and visual inputs that were manually checked by the experimenter before the robot initiated the holiday planning task. During the interactions, the experimenter listened in on the conversation using earphones in order to manually provide the system with the participant's preferences and perform additional control actions when needed (Wizard of Oz).

After the final dialog phase, the software was automatically terminated after which the experimenter returned to the participants to remove the headsets and hand back the questionnaires. Participants were instructed to answer the remaining questions on their own, and to call for the experiment leader if any assistance was needed. Participants spend on average 10 minutes to complete the questionnaire while the experimenter receded in the control room. After completion, the questionnaires were collected. Participants were verbally debriefed and informed about the exact purpose of the study and provided answers to any remaining questions. After a word of gratitude for participation, reimbursements were handed out. The participants were then reminded not to forget any belongings and escorted out of the lab room. Any valuable feedback received during debriefing was written down.

Participants

Multiple turn takes initiated by the robot created the opportunity to do multiple measurements during each condition. Based on the theoretical framework for turn taking by Heldner et al. (2011) and (Bohus & Horvitz, 2010, 2011a, 2011b) an average of 49 speaker-turns (i.e. measurements) were expected to occur in twenty-minute lasting dialogs with the robot. In related research of Trafton et al. (2008), the perceived naturalness of a robot is compared between two conversation-behavior models that vary by 500ms in reaction-speed timings. Given similar timing manipulations between conditions for the current study, the effect size of Trafton (*f ≈ .55)* was used to calculate the required sample size and predicted power. According to the specification in Table 2 below, approximately 343 (7*49) qualitative measurements per factor for 20 minute lasting dialogs could be expected.

Table 2: Pre-experiment power analysis

ANOVA - ANALYSIS INPUT		***OUTPUT SPECIFICATION (G*POWER SOFTWARE)***	
Effect size f	*= .55*	Non-centrality parameter λ	= 11.89
α err prob.	*= .05*	Critical F	= 3.26
Power	*= .85*	Denominator df.	= 36
Numerator df.	*= 2*	Minimal sample size	= 39
Number of groups	*= 3*	Participants per factor[1]	= 14
		Conversations per factor	= 7
		Required sample size	= 42
		Predicted power	= .88

To recruit participants, flyers as well as invitations by email have been used to encourage people to participate in the study (see APPENDIX H). All subscriptions were managed by the JF Schouten (JFS) School database for User Interaction System Research, which is part of the Human-Technology Interaction (HTI) department of the Eindhoven University of Technology (TU/e). As required, a total of 42 participants took part in the experiment. Arriving in duos, the participant pairs were randomly assigned to one of the three experiment conditions. The demographics comprised of 20 females and 22 males (mean age 25.10, SD = 5.24, range 19 to 47), of which only 2 participants were native English speakers, 26 native Dutch, complimented with persons from a large variety of 9 other mother tongues. From the total of 42 participants, 15 of them (35.7%) regularly used a speech interface, of which 'Siri' from Apple Inc. is the most mentioned voice assistant (7x), followed by 'Google Now' (4x). Nearly half (22) of the participants had interacted with a (similar) robot before. As reimbursement for partaking in the experiment, participants received a fee based on fixed rates for either students or non-students.

[1] The rounded up value of the minimal sample size, divided by number of factors

MATERIALS

Hardware

To realize the interactive verbal robot, several pieces of hardware besides the Nao robot were used; a standard wired network router, a pair of wireless microphone headsets with receivers (Sennheiser EW100 G3), a 3.5mm double-mono to stereo adapter, and a laptop with line-in audio port running MS Windows 7. Scripting is done in the Python 2.7 programming language, using several build-in libraries of the Python(x,y) distribution and additional open source libraries such as OpenCV and PyAudio, in combination with the Naoqi API (Aldebaran Robotics).

The system was roughly be divided into two parts; a combination of underlying subroutines and a behavioral state machine (see graphical representation in APPENDIX E). The subroutines processed raw data provided by the aural and visual sensors, and converted it to information needed to regulate transitions in the behavioral state machine (for data acquisition details see APPENDIX F). An overview of the implemented subroutines will be provided first, followed by a description the state machine.

Software

Audio sampling

A big challenge when using the Nao's built-in microphones, was to realize reliable speech recognition and localization, especially during moments of overlapping speech when both participants speak. A simple workaround for this problem was using individual head microphones to determine the source of speech. This was less complex and less prone to errors compared to using acoustic localization algorithms in combination with the Nao's built-in microphones, or the need for external microphone arrays. Using the PyAudio library, an audio sample rate of 44.1 kHz, and a block time of 20ms, the Root Mean Square (RMS) value was calculated for the two individual audio channels (split stereo) representing the output power of the participants' microphones. Using basic filtering methods to partially compensate for background noise and cross-channel interference, a threshold was defined to differentiate between the presence and absence of speech. Additional features were implemented to filter out RMS-peaks and to balance out threshold sensitivity in relation to speech-activity responsiveness.

Video sampling and keystroke capture

A limited but still usable framerate of 10 frames per seconds (fps) on average was achieved for the Nao robot built-in camera. This was done by means of a dedicated router with a cabled network connection between the Nao and the computer system, proper light conditions, and in addition by using grayscale capture settings at a mediocre VGA resolution (640x480). Using OpenCV, the 'Haar Frontal-face-feature Cascaded Classifier' face detection algorithm is used to identify the presence of a single face. Several performance measures were modified, such as the detection robustness and adjustment speed of the variable face-search region. After successful detection, the neural-network-based head pose estimator developed by D. van der Pol (2010) provided an estimation of the yaw and pitch of the face relative to the camera's point of view. Similar to the implementation of Klotz et al. (2011) and Johansson et al. (2013), the image from the Nao robot's camera was first used to detect people's faces and subsequently to estimate their current visual focus of attention. After calibration the yaw angle achieved ±0.2 degrees precision within the horizontal operating range stretching from -25 up to +25 degrees. Due to the limited operation range, only a reliable dichotomous estimation was used to discriminate a participant that was focusing towards the robot or in contrast looked away from it. During moments of face detection failure, the most recent face coordinates were temporally used in order to make new head pose estimations. This slightly increased the robustness of the visual system, and was only applicable due to the fixed positions of participants in the experiment. Subsequently, the running average of 5 head pose estimations was used to smoothen the output. Using the default OpenCV 'WaitKey' function, manual key presses could be detected and were used to realize experimenter control during runtime.

Speech-to-Text import

In order to not only detect participant speech but also be able to capture and recognize what words are being said, a subroutine was created that enabled to import results from commercially available speech recognition software. For this study Nuance Dragon Naturally Speaking (DNS) was used during the development. With help of the DNS Administrator Guide document, advanced settings could be accessed to properly setup the software. The recognition software used a settings profile that provided the fastest possible (partial) translation results, and enabled immediate logging of recognition results. Using Python's 'linecache' function, a subroutine continuously kept track of any chances in the recognition log, and imported the newest recognition results as a list of words. The text-to-speech subroutine was built with the intention to include a shallow semantic parser to the system to derive meaning from participants' utterances[2]. When an adequate level of understanding would be realized, autonomous dialog responses would become feasible (low level artificial intelligence). However for the current study, unavoidable recognition errors as well as imperfect semantics posed a too big risk. Despite promising results that would suffice for less critical verbal interaction purposes, the text-to-speech subroutine was not used and was 'replaced' by a Wizard of Oz implementation. This required a human controller to personally interpret utterances from participants and manually provide the system with the necessary information.

SUB-process input handler

The combination of processes just described basically formed the eyes and ears of the robot. To realize certain behavioral characteristics the combination of all processes needed to run reliable, be easy accessible, as well as fast and efficient. Since the subroutines were independent from each other and used different sources, the use of parallel processing (multithreading) provided an ideal solution to achieve fast system responsiveness. Initiating, maintaining, and properly terminating subscripts was the main function of the SUB-process input handler. Using queues, it made sure to collect only the most recent available data entries. Information from all subroutines were combined into a single data class. This class was returned upon request to the behavioral state machine (see APPENDIX E for details about the data class).

Dialog manager

A critical part in the development of the verbally interactive robot, were the dialog capabilities that were determined by context and richness of utterances available to the system. For the current study a dialog schematic was implemented with the goal to elicit and maintain a conversation for at least 10 to 20 minutes regarding a holiday planning task. An overview of the constructs that were defined, which provide insight in the way the dialog manager works, is included in APPENDIX C. More generally, the custom build dialog manager used predefined ordered dialog-stages that were processed one by one. During each dialog stage (11 in the current study) the robot asked the participants a question and provided them with a set of answer options.

[2] URL to semantic parser example:
http://demo.ark.cs.cmu.edu/parse?sentence=so%20what%20do%20you%20think%20of%20me

To maintain and exert control over the speed of the conversation, the robot asked for the opinion of each participant. Hereby it sometimes referred to a specific option, as well as asked to elaborate on one's own argument, or comment the response of the other participant. In some cases the robot provided specific information about one of the answer options, or reminded the participants of the available set of choices. The number of times these type of reciprocal questions were initiated by the dialog manager, were by default equal to the number of preprogrammed answer possibilities. However, the number of questions asked as well as the order of different preprogrammed questions, could be modified. In addition, a feature was implemented that allowed a controller to skip any remaining questions of a dialog phase during the experiment (i.e. speed-up function). This function was particularly useful in situations when two participants were more or less finished talking, to avoid redundant questioning.

After all pre-selected questions had been asked, or when a speed-up took place, the dialog manager constructed a verdict utterance. This verdict was meant as feedback to the participants, reflecting to both of them the understanding of the robot regarding their preferred option of choice. Although the verdict depended on correct input from the Wizard, or some automated speech processing module, the dialog manager did not depend on it. Without the need for waiting on input, the dialog manager's construction ruled out any external influences on the responsiveness of the system, and ensured the continuation of the dialog. However, most utterances in subsequent dialog phases did depend on previous phases (e.g. preferred country determines city suggestions), which led to the following 4 deterministic verdicts:

- *When the preferable option is unknown by the system*; the dialog manager (DM) will select a 'verdict utterance' that expresses the system's uncertainty about the participants' preference and selects randomly one of the options as definite answer.
- *When none of the options are preferred by the participants*; the DM will render a verdict that will let the participants know that; none of the suggested options seem to fit their criteria and hence an alternative option is chosen that was not mentioned before to the participants.
- *When multiple options are preferred*; the DM provides an utterance that informs the participants about the multitude of suitable options and will select one of them for simplicity (at random).
- *When a single option is preferred by the participants*; the DM will acknowledge the apparently successful discussion to reach mutual agreement and selects the desired answer in order to continue.

After informing the participants which option is selected, the cycle repeated for the remaining consecutive dialog phases. For an utterance to be directed towards a specific participant, the dialog manager was able to embed behavioral commands in string literals. This way certain actions could be initiated that had to accompany the utterance being pronounced (e.g. a head turn). Despite all of the above mentioned constructs that formed the backbone of the dialog capabilities of the robot, typical systematic repetitiveness was likely to become very salient to participants. This would negatively influence the verbal interaction experience. To counteract this side effect, an abundance of utterances have been defined. Many utterances consisted of mutable content allowing for a multitude of textually different sentences that could be pronounced by the robot. This corresponded to the first rule in a set of heuristic rules defined by Csapo et al. (2012) stating that the robot should "not give the same instructions to the user in the same way over and over again" in order to create more interesting dialogs between a person and the robot. Due to the experimental nature and focus of this study, no additional considerations were made to conform to the second heuristic that requires the robot to "vary the level of sophistication in terms of functionalities". In APPENDIX D a collection is provided with the specific text of all possible robot utterances.

Since the system was not processing anything of what was actually being said, it was the wizard's responsibility to provide the system with correct information, to ensure a correct flow of the dialog content. Important to realize is that besides minimal control over the number of sentences pronounced by the robot, no influence what so ever could have been exerted on the turn take timing of utterances. Johansson et al. (2013) implemented the possibility for a wizard to exert control over the content of robot utterances, as well as when this content was verbalized. In the paper it was explicitly concluded that the human-robot head pose patterns that were compared to human-human patterns had likely been "affected by the turn taking strategy employed by the wizard". To rule out any similar influence and experiment confound, the control options for the wizard were kept to a minimum.

Behavioral state machine

For flexibility and implementation convenience, a finite state machine design was chosen to realize the robot's behavioral capabilities. Being the center component of the total system, it relied on the previously discussed components that provided the information necessary to make higher level inferences concerning the robot's environment. The robot's behavior was defined by a deterministic model. Transitions between internal states as well as the implementation of turn take strategies themselves, were defined by rules. These rules mainly comprised of Boolean tests on input parameters from sensors that provided information about the robot's environment. Data relevant to the system to make inferences was assumed to be reliable enough to act upon at any point in time. Using these inferences, appropriate system actions could be triggered that often translated to low level behavioral expressions like initiating movements, controlling leds, or producing artificial speech. The state machine for this study consists of 7 individual states and 5 conditional transitions, as illustrated below in Figure 5.

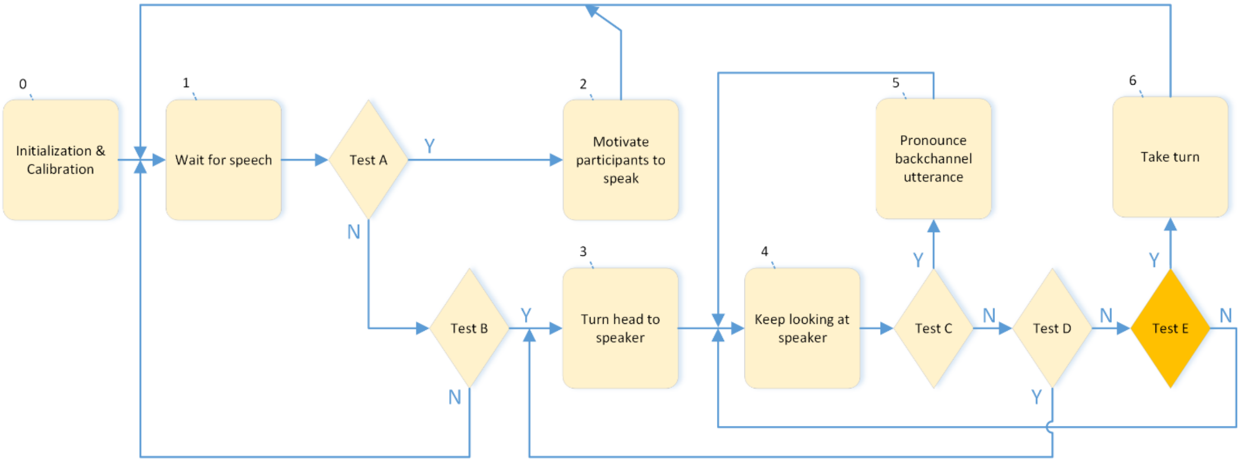


Figure 5: Robot behavior diagram

0. *Initialization & Calibration*; in this state the subroutine functions were initiated and connection was made with the Nao robot. After successful startup the robot was able to turn its head, simulate eye blinking with leds surrounding its eyes, indicate the source direction of detected speech using reactive leds surrounding the robot's 'ears', and produce speech. After introducing itself as 'Marvin' and welcoming the participants, Marvin asked the participants one after the other to introduce themselves. Emphasized by Mutlu, Shiwa, Kanda, Ishiguro, and Hagita (2009), the opening of a conversation by means of greetings is an important activity since speakers signal their conversation

role intentions and likewise communicate their availability to any role imposed by others. Also Goffman (1955) describes greetings as serving to “clarify and fix the roles that participants will take during the occasion of the talk and to commit participants to these roles.” Without further notice to the participants, besides the instruction to clearly look at the robot while they introduced themselves, the system used the camera stream during these introductions to calibrate the head pose estimator. Manual input of the controller was required to redo a calibration or confirm the success of an attempt. Several automatic re-calibration mechanisms were also implemented to ensure proper calibration. After the introductions, Marvin informed the participants about the experiment and asked each participant explicitly if everything was understood before beginning the experiment task. After confirmation the experiment leader enabled the system to continue to state 1.

1. *Wait for speech*; arriving at this state, either after startup or after a previous system cycle, the robot asked the participants a holiday planning question with associated answer possibilities that were to be discussed. At the beginning of the question asked by the robot, a behavior was implemented based on the animation of the interactive avatar by Bohus and Horvitz (2010). After having looked at both participants at the beginning of the turn, the Nao robot adopted a neutral view while finishing its utterance. Not looking to any of the participants at the end of a turn was to stimulate them to start a mutual discussion first about the new topic before directing themselves to the robot.
   - *A.* Subsequently the system evaluated test A (conditional transition). It checks if no participant speech was detected for more than 4 seconds. If this is the case, it assumed the participants were having trouble to get a discussion started and goes to state 2.
   - *B.* When a prolonged silence was however not (yet) detected, test B was evaluated to determine if one of the participants had already started to speak for more than 500ms. If not true the system stayed in state 1, otherwise state 3 became active.

2. *Motivate participants to speak*; the robot tried to motivate participants to start a conversation, by briefly looking at them and ask who is willing to speak up. The exact used sentence was determined by the dialog manager (DM) that provided the state machine in this case with a ‘breaksilence’ type of utterance (see also APPENDIX C and APPENDIX D). After this action the system returned to state 1.

3. *Turn head to speaker*; the robot initiated a head turn to face the speaker and moved to state 4.

4. *Keep looking at speaker*; in this state the robot’s head will tracked the face of the talking participant. Due to subtle movements when adjusting the head to direct its reciprocal gaze towards the speaker, the illusion was enforced that the robot actively listened to what was being said.
   - *C.* While gazing towards the speaker, test C was evaluated to detect if the speaker was already talking uninterrupted for more than 4 seconds. If this was the case, the system moved to state 5.
   - *D.* If this was not the case, the system checked if there had been a change of speaker. Test D evaluated if uninterrupted speech was received for more than 500ms from a speaker opposite to the currently looked-at participant. If this was true, the system moved back to state 3. This way the robot noticeably kept track of the conversation flow. This was similar to the implementation by Trafton et al. (2008). They explored among other things how natural naive participants would perceive a robot that changes its visual attention in a conversation to a new speaker after a half second delay. Informal tests revealed that this gazing behavior was perceived quite natural.

E. When however none of the two tests above resulted in a state transition, dialog conditions were fit for the robot to test whether or not it should take turn. Test E was the final and most important transition for the purpose of this study. Depending on the active experiment condition, test E represented one of the three turn take strategies that have been defined in the methodology. In short, besides a gaze cue the conservative condition required 1 second of silence before the robot took turn, in contrast to the assertive condition that only needed an instant of silence from the current speaker. The adaptive condition allowed the robot to take turn after a non-fixed delay, which could cause overlap with the speech of a talking participant. Details about the implementation of the turn take strategies will be described in the next paragraph. When the active strategy (test E) resulted in a turn take, the actual turn was carried out in state 6.

5. *Pronounce backchannel utterance*; in this state the robot pronounced a short backchannel utterance like “I see” as a sign of interest and signal the talking participant that it is still paying attention. The utterance was defined in the dialog manager as ‘longturnindicator’ and had a few varieties that can be looked up in APPENDIX C & D. This mechanism also prevented the robot from being completely muted whenever a participant speaks for a longer period of time. After this action, the systems returned back to state 4.

6. *Take turn*; this state executed the actual turn take. The dialog manager determined the type of utterance to be pronounced (and thereby its content) and to which participant the utterance had to be directed (i.e. initiating a head turn). When the robot had finished its turn and stopped speaking, the behavioral cycle was complete and the system returned to state 1.

Modeling turn take strategies

Conservative condition

Just as the experimental turn take models are different from each other, the implementation of the conditions also differed substantially. Starting with the conservative condition, it was unknown when a turn take gaze cue was to be expected relative to the start of a silence cue. Therefore a division was made between two scenarios; 1) either a gaze cue is received before the start of a silence cue, or 2) a gaze cue is received after the start of a silence cue during the silence period itself.

Silences during a conversation occur frequently with varying durations that are not known in advance. When a period of non-speech can be considered a genuine turn take silence cue, depends on the self-defined minimum required silence-time duration. This duration needs to be balanced to minimize the chance of cut-ins but still be quick enough to maintain adequate system responsiveness. Similar time constraints are however not typically defined for gaze cues. To allow for flexibility in timing occurrences of the two cues relative to each other, the concept of a ‘turn take opportunity window’ (interval) has been implemented. With the exception of the adaptive condition, which will be covered later on, the turn take window defined the interval in which both cues combined had to be present in order for the robot to execute a turn take. The window stretched over a predefined time range and was initiated by the cue that was detected first. Two kinds of windows could therefore be described; the ‘gaze before silence’ window and the ‘silence before gaze’ window.

Both serve the same purpose and are similar to each other, with exception on the rule(s) that will cause the windows to get de-activated. The gaze before silence window will only get deactivated when the window duration time has passed without any new gaze occurrence, and gets (re-)activated upon any instant of gaze. The silence before gaze window gets deactivated when the predefined window duration time has passed or when the silence cue is violated during the active window time. It gets activated only after a predefined time period of non-speech (i.e. silence) has passed. When during an active window the opposite cue to the window-activation cue is detected as well, a turn take is initiated by the system.

For the current study the duration of the turn take opportunity window was set for lasting maximally 2 seconds when triggered by a silence cue, while lasting 2 seconds + the required silence cue duration when triggered by a gaze cue. Any occurrence of absence of speech equal to 1 second or longer is seen as a genuine turn take silence cue for the conservative condition.

The turn take strategy for the described conservative condition is illustrated in Figure 6, showing the two possible ‘window scenarios’ that lead to turn takes.

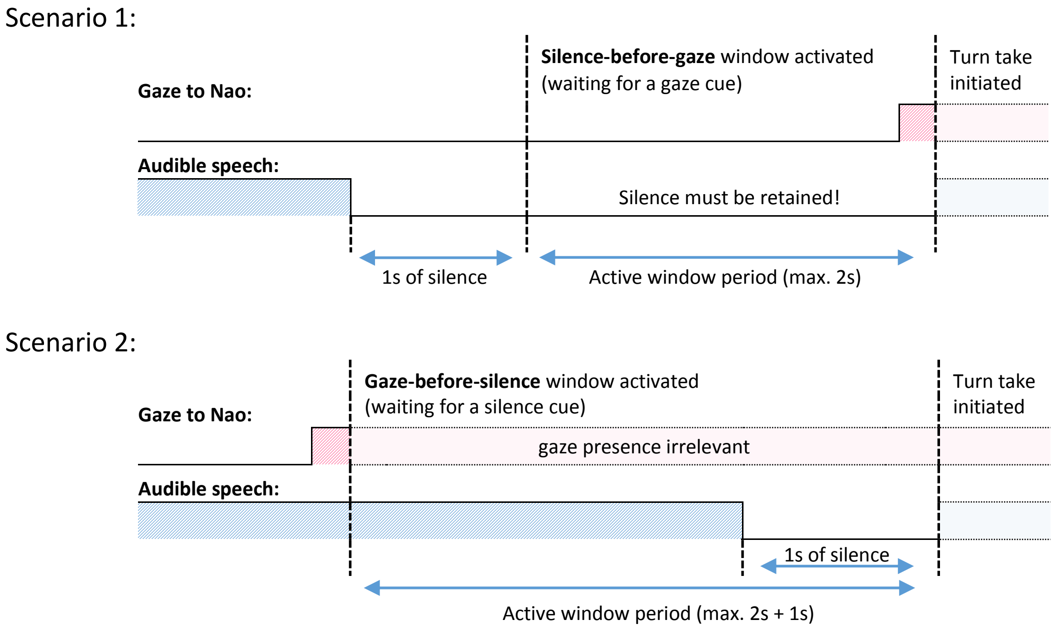

Scenario 1:
Silence-before-gaze window activated
(waiting for a gaze cue)
Turn take
initiated
Gaze to Nao:
Audible speech:
Silence must be retained!
1s of silence
Active window period (max. 2s)
Scenario 2:
Gaze-before-silence window activated
(waiting for a silence cue)
Turn take
initiated
Gaze to Nao:
gaze presence irrelevant
Audible speech:
1s of silence
Active window period (max. 2s + 1s)

Figure 6: Conservative condition[3]

[3] Graphical representations not drawn to scale

Assertive condition

The same windowing principle was used for the assertive condition of the experiment. The goal of this condition is to implement the fastest possible system response. Therefor the required silence duration was reduced to last only an instant, just like the gaze cue. Due to the lack of a silence duration requirement, turn takes were very likely to be initiated too soon resulting in higher cut-in rates. Hence the validity of the silence cue was compromised which made the order of cue appearance irrelevant as well as the need for two distinctive window 'types'. This being the only difference compared to the conservative condition, resulted in simplified scenarios for the assertive condition that used a single generic turn take opportunity window as illustrated below in Figure 7.

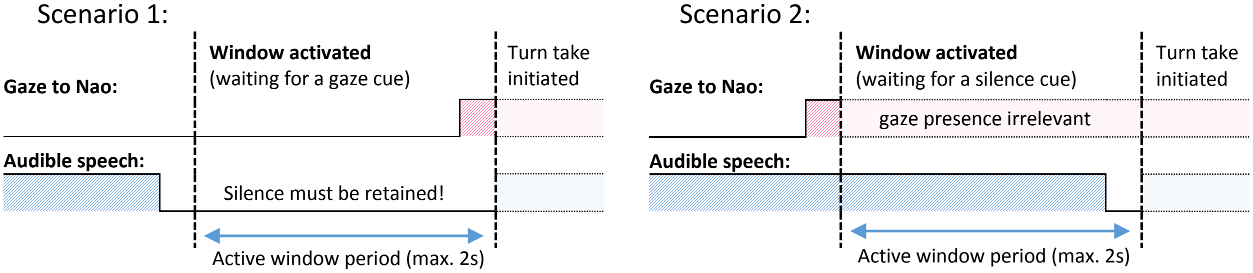


Figure 7: Assertive condition

Adaptive condition

The remaining turn take model to be defined is the experimental adaptive condition. Instead of predefined timing parameters, the adaptive condition was designed as a self-correcting system with the goal to identify an 'optimal' delay between the occurrences of the gaze cue relative to the silence cue. Adjustments to this delay were made in an effort to increase the system's turn take proficiency, thereby minimizing cut-ins and maximizing response times.

Like the other experimental conditions, the gaze cue was a prerequisite for any turn take action. Assuming that a gaze cue did not need to last a certain period of time, any delay before an actual turn take was counted from the first moment of gaze. This resulted in a minimum possible delay of zero seconds which would lead to a turn take executed immediately after the occurrence of a gaze cue. To maintain experimental consistency, the maximum possible delay was defined to be 3 seconds, which corresponded to the maximum time interval between the two cues in the conservative condition (2s window duration + 1s silence).

Unlike the other conditions however, the absence of speech was not a prerequisite for a turn take. This effectively allowed the system to make speech overlapping cut-ins, which was generally speaking undesirable behavior. For the current condition however, a cut-in event provided valuable information used by the system to correct turn take timings. In order to determine what kind of correction is needed, the presence or absence of speech at the moment of a turn take was measured, as well as the absence of speech at the moment when a gaze cue was received.

The gaze cue triggered measurement of speech absence, referred to as the pre-silence variable, combined with the information of a cut-in occurrence, resulted in two variables with four possible scenarios. Shown in Figure 8, these scenarios were used to inform the system about its current turn take proficiency.

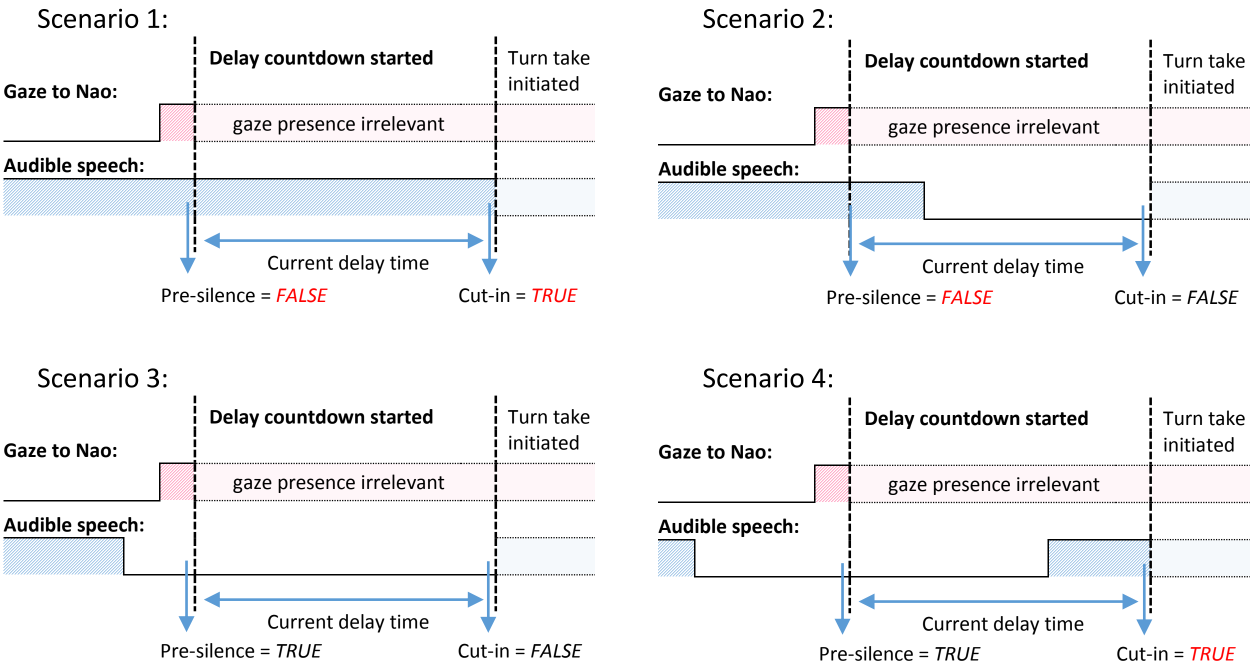


Figure 8: Adaptive condition[4]

Operating within the boundaries of the earlier mentioned minima and maxima delay values, these four scenarios were used to define turn take delay-adaptation rules. The delay-adaption options were dichotomous, either the current delay for a turn take after a gaze cue had to be prolonged or shortened. The default delay-before-turn take was relatively short lasting 300ms, a value that allowed for a responsive system right from the start. For scenario 1 and 2 the pre-silence variable is *FALSE* and the subsequent cut-in variable is either *TRUE* or *FALSE*. Expecting to find a continuous silence period suitable for a turn take, the first scenario depicts a situation that required the system to increase the delay time since no silence was present both before and at the moment of turn take. For scenario two the system did not make a cut-in, which is preferable but could still require the system to increase its responsiveness and therefore shorten its current delay time.

For scenario 3 and 4 the pre-silence variable is *TRUE*. When the current settings result in a proper turn take without cut-in as depicted in scenario three, the system was instructed to find the maximum possible delay time by increasing the delay. Although this rule may seem counterintuitive, it is able to provide useful timing information concerning the maximum silence duration measured from the moment of a received gaze cue. In addition, the rule contrasts the delay adaption rule used for the fourth scenario. In the fourth scenario the system's turn take resulted in a cut-in. Given the presence of silence at the moment of gaze, the current delay had to be reduced in order to prevent cut-ins in subsequent turn takes.

[4] Graphical representations not drawn to scale

An overview in Table 3 of the rules just described:

Table 3: Delay adaptation rules

	Turn take scenarios				***Delay adaption rule***
1	Pre-silence = FALSE	&	Cut-in = TRUE	→	Delay is too short and must be increased (+)
2	Pre-silence = FALSE	&	Cut-in = FALSE	→	Delay might be too long and is shortened (-)
3	Pre-silence = TRUE	&	Cut-in = FALSE	→	Delay might be too short and is increased (+)
4	Pre-silence = TRUE	&	Cut-in = TRUE	→	Delay is too long and must be shortened (-)

Staircase procedure

What remains to be defined is the magnitude of changes applied to turn take delay times. In order for the system to find a potential sweet spot in the delay-before-turn take, the number of possible adaptations and the used timing resolution were considered the two most critical factors for success. The number of adaptations were equal to the number of turn takes made by the system, which was unknown beforehand and was likely to vary between experiment sessions. In addition the resolution of timing adjustments depended on the total time range to be covered, defined by the difference between the minimum and maximum allowed delay which turned out to be a relatively large range of 0 (no delay) to 3 seconds.

Using a fixed step size (e.g. ±200ms) was considered a suboptimal implementation for delay timing adjustments. Such a system would be a compromise between mediocre timing precision and feasible number of turn takes necessarily to effectively operate over the entire 3s range. Instead, preference was given to a variable step size implementation based on a staircase procedure. Given the audio sampling block length of 20ms, the smallest possible step size was likewise 20ms. Starting with an arbitrary chosen default step size of 100ms, the largest possible step size was set proportionally at 500ms. Every time a turn take was initiated by the system, resulting in a plus (+) or minus (–) delay adaptation rule, the corresponding change in step size was either; equal, half, or double the previously used step size. Which step size was applied depended on the history of the previous 3 subsequent delay adaptation rules. Whenever the same delay increment was needed three times in a row, the next increment doubled the step size leading to an increment that was twice as large. The inverse rule applied whenever three times in a row the same delay decrease occurred, leading to a double decrease when the same rule was applied for a fourth time. However when the previous three adaptation rules occurred in alternating order using the same step size (i.e. with alternating sign), the timing resolution needed to be improved by halving the step size. This allowed the system to 'zoom-in' on any potential sweet spot. Any other order of earlier applied adaptation rules resulted in an unaltered step size. A table overview of the staircase procedure just described:

Table 4: Staircase procedure

	Rule occurrence order				***Step size change***
1	+ + + +	or	− − − −	→	Step size is doubled
2	+ − + −	or	− + − +	→	Step size is halved
3	Remaining (arbitrary)			→	Step size remains unaltered

System log

To ensure data retrieval for post-experiment statistical and time sequence analyses, a logging function stored the most informative variables during run-time. Having already explained most of them in this chapter, an overview of all the stored variables is found in Table 5.

Table 5: Log file contents

Variable name		***Description***
Experiment number	=	The session number of the current experiment
Experiment condition	=	The type (number) of experimental manipulation (conservative = 1, adaptive = 2, assertive = 3)
Turn take window	=	The set duration (s) of the system's turn take opportunity window
Timestamp	=	Time of log entry (ms), needed due to variable time intervals between log entries as result of varying processing delays
Iteration	=	Number of passed sensor data retrievals
Mic input	=	Current speech activity (indicating silence or who is speaking)
Face detected	=	Boolean value that becomes true when a participant's face is detected
Calibrated yaw	=	The positive or negative yaw angle of the detected face, used as an estimate for participants' gaze directions
TTS words	=	The words that are detected by the speech recognition software
Nao utterance	=	The words that are pronounced by the robot upon a turn take
Nao gaze direction	=	The gaze direction of the robot, either left, right, or neutral
System behavioral state	=	The current active state (number) of the 6-state behavior machine
Dialog phase	=	The current phase (number) of the dialog manager sequence
Take turn	=	Boolean that is only true for a single instant when a turn take is done
Gaze before silence & Silence before gaze	=	This is a unique variable that only applies to the conservative condition, and reports which of the two cues in respect to each other triggered the turn take opportunity window, that eventually led to a turn take
Pre-silence & Cut-in detected	=	This is a unique variable that only applies to the adaptive condition, and reports the state of the two cues around the moment of turn take. The four possible combinations effect the system's subsequent turn takings
Fixed/Adaptive delay	=	The user defined or system defined delay (s) before turn take

Additional system specifics and features

As briefly mentioned before, in order to create and enhance the illusion of liveliness and attention, the leds surrounding the robot's eyes were programmed to mimic the occurrence of natural eye blinks. In the meanwhile the robot's build-in face tracker function made sure the Nao's gaze followed the head movements of the participant it looked at. Together these functions prevent the robot's head to become static when turned in the direction of one of the participants. Another functionality enabled the robot to provide subtle feedback to the participants to indicate that it was detecting speech. For this purpose the robot's 'ear' leds lit up blue on the side that corresponded to the sound's origin, which was either the participant located on the left or right side of the Nao. Also, in order for the robot to more properly match the role as conversational moderator and its sincere intension to take part in the group dialog, the voice of the Nao was somewhat altered (e.g. lower pitch). Some features have been implemented to also simplify the development, with the goal to increase general purpose employability, as well as enhance the usability of the framework to control the Nao robot directly or run simulations. Although not visualized in APPENDIX E, these are some of the features:

- Unconditional switching from built-in camera stream to any external camera stream.
- The option to record and store microphone audio and camera video data (separately).
- The possibility to rapidly switch between multiple (built-in) face detection algorithms.
- Nao independent design; automatically converting Nao behavioral commands to console text when not connected, and using PttsX to synthesize robot utterances to audible speech which allows to run almost fully functional simulations of any behavioral state machine.

RESULTS

Preliminary Data Analysis

Design and power

The number of measurements ended up to be lower and also variable for each condition (*M* = 256, *SD* = 41) compared to the predicted number of turn take measurements. This was caused by shorter conversation times and differences in dialog lengths across conditions (*M* = 605.30 s, *SD* = 179.88 s). Twenty minute lasting dialogs could not be achieved due to the disproportional extra time needed to prolong the dialog capabilities of the robot. The decline in number of robot turns for in particular the conservative condition, is the result of the 'speed-up' command (Wizard of Oz control). Although the robot was programmed to make a fixed number of turns during each dialog stage, the speed-up option was used whenever remaining turns became obsolete and to maintain progression in the dialog flow. This option was needed more for conversations in the conservative condition, because participants' turns lasted long enough for them to thoroughly discuss matters before the robot took turn again. Without a speed-up option the robot would unnecessary provide extra information or ask follow up questions to answers that participants already finished discussing about. This would make it very obvious to the participants that the robot is not listening at all. The average number of turns and additional turn taking information per experimental conditions is shown in Table 6.

Table 6: Dialog characteristics per condition

Conditions	*Total number of turn take measurements*	*Average dialogue duration (s)*	*Average robot turn take occurrences*	*Average number of robot turns for 10 minute dyads*
Conservative	237	*M* = 796.39, *SD* = 179.88	*M* = 34, *SD* = 3	26
Adaptive	229	*M* = 506.85, *SD* = 26.15	*M* = 39, *SD* = 3	46
Assertive	303	*M* = 512.68, *SD* = 49.21	*M* = 43, *SD* = 2	51

Since more than one factor of interest is measured, a combination of significant effect sizes enabled a recalculation of the power for the current experiment. Analysis revealed an average achieved power of .79 which is slightly lower than the predicted power of .88 as mentioned in the research design. This is however still adequate enough given Cohen's (1988) benchmark of 0.8, especially since four out of six factors lie above this threshold as can be observed in Table 7.

Table 7: Post-experiment power analysis

Factors	Main effect size (η^2)	*N* (*df* = 2)	Power (Avg. = .79)
SASSI response accuracy	.504	42	.804
SASSI likeability	.482	42	.765
SASSI annoyance	.566	42	.890
SASSI speed	.524	41	.824
Godspeed likeability	.523	42	.834
Godspeed perceived intelligence	.418	39	.597

Data screening

The first step of data filtering required the elimination of noise from very short (meaningless) utterances. The minimum useful utterance duration is determined by the pronunciation time required to pronounce affirmations or contradictions like “yes”, “no”, or verbally similar short utterances. By means of visually inspecting waveforms from recordings of the aforementioned short utterances, the minimum pronunciation time at nominal speed was set to 250ms. Consequently all utterances shorter than 250ms have been eliminated from analysis. Consequently most turn overlaps, a short utterance that overlaps within a (much) longer uninterrupted utterance of another participant, have been eliminated from the data as desired. Remaining turn overlaps (>250ms) have subsequently been filtered out. To parameterize the speech of participants, all utterances that were cut-off by interrupting turn takes of the robot have been excluded from analyses. For a fair group conversation, the turn take models were designed to provide participants equal chances to speak by giving them the floor when the robot yielded its turn. The data shows there was indeed a good balance, since 53% of the time the participant seated on the left started to speak first, versus 47% for the participant on the right.

To assess the successfulness of the three predefined turn take models, a comparison is done between the intended turn take delay times and the Real Turn Take Delays (RTTD). RTTD is the duration of silence that was present between the end of a participant’s turn and the start of the robot’s turn. In Figure 9 the RTTD is visualized for each of the three experiment conditions.

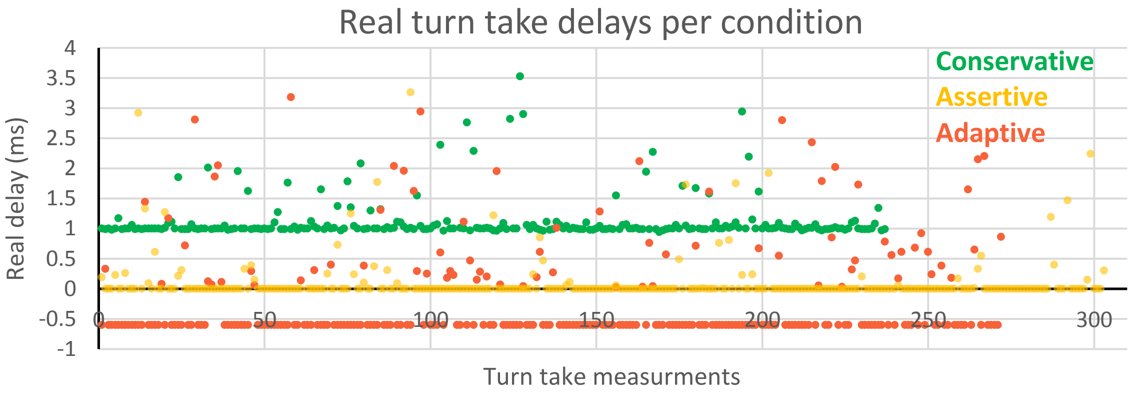


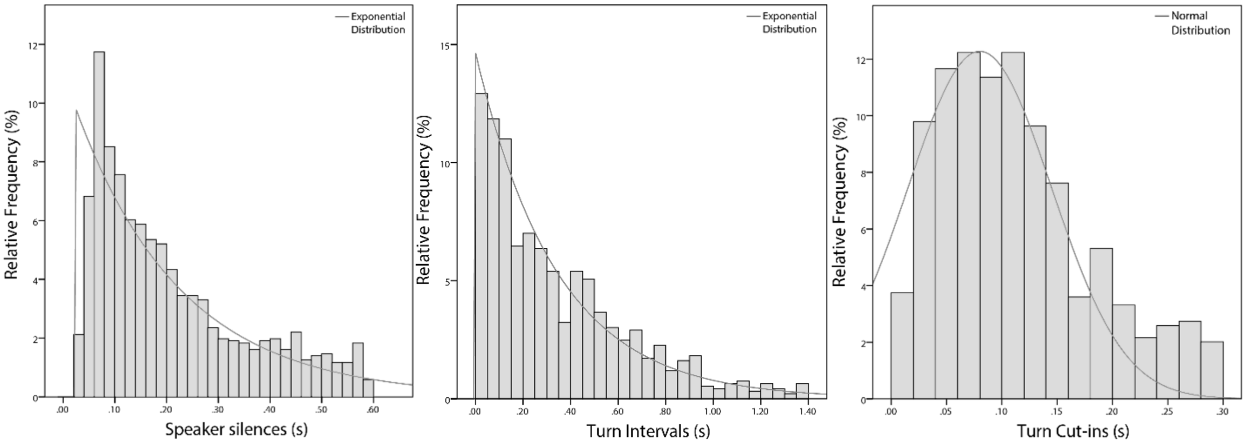
Figure 9: Real delay times per condition [5]

Apart from the adaptive condition, turn take model success rates can be calculated. Compared to the predefined required silence duration of 1 second for the conservative condition and ≈ 0 seconds for the assertive condition, there are several measurements that exceed these thresholds. Excessive turn take delays can be caused by moments of inadequate detection of gaze or absence of speech. Timing analysis of cues for these turn take measurements can therefore pollute the data from cue timings retrieved from successful measurements, and are marked as outliers (i.e. excluded from analyses). 15% (36 out 237) of the conservative samples >= 1.12s were marked as outliers, which yields 85% successful turn takes for the conservative condition. The outlier threshold was not exactly equal to 1s due to sample frequency differences in audio and gaze-cue measurements. In a similar way for the assertive condition, 17% (51 out 303) of the assertive samples >= 0s were considered an outliers, which results in 83% successful assertive turn takes. The negative measurements represent cut-in occurrences in the adaptive condition.

[5] Negative sample values are arbitrary but are used to indicate cut-ins

Dialog Analysis

Speaker silences and turn intervals show clear exponential distributions. The distribution of turn cut-ins seems however to follow a positively skewed normal distribution. Figure 10 depicts the different distributions and corresponding descriptives.



	Speaker silences (s)	*Turn Intervals (s)*	*Turn Cut-ins (s)*
Sample size	*N* = 1363	*N* = 928	*N* = 695
Mean (SD)	0.20 (0.15)	0.34 (0.31)	0.11 (0.07)
Left-sided 95% CI	0 – 0.55	0 – 0.95	0 – 0.25

Figure 10: Speech histograms and descriptives

To test for differences in distribution for the variables between the three experimental conditions, the non-parametric Kruskal-Wallis test is used. A comparison between experimental conditions revealed no significant differences for speaker silence durations, $H(2) = 4.65$, $p = .10$. Also no differences were found for the duration of turn intervals, $H(2) = 4.978$, $p = .08$, or turn cut-ins, $H(2) = 1.565$, $p = .457$. Between the three conditions, a significant difference was however found for the utterance durations of participants, $H(2) = 20.08$, $p <.001$.

Multiple post-hoc analysis using the non-parametric Mann-Whitney test revealed that the utterance durations in the adaptive condition (*Mdn* =.78) significantly differed from both the conservative condition, (*Mdn* = 0.95, *U* = 421137.5, *Z* = -3.98, *p* <.001, *r* = - .09), and assertive condition, (*Mdn* = 1.01, *U* = 149416.5, *Z* = -3.96, *p* <.001, *r* = - .12). Shorter utterances occurred more frequently for the adaptive condition, while longer utterances were less frequent. No such difference is found between the conservative and assertive condition (*U* = 416775.5, *Z* = -.957, *p* =.339, *r* = N/A). In Figure 11, the similar distributions of the conservative and assertive condition are combined on the left side, which enables comparison with the distribution of the adaptive condition on the right side.

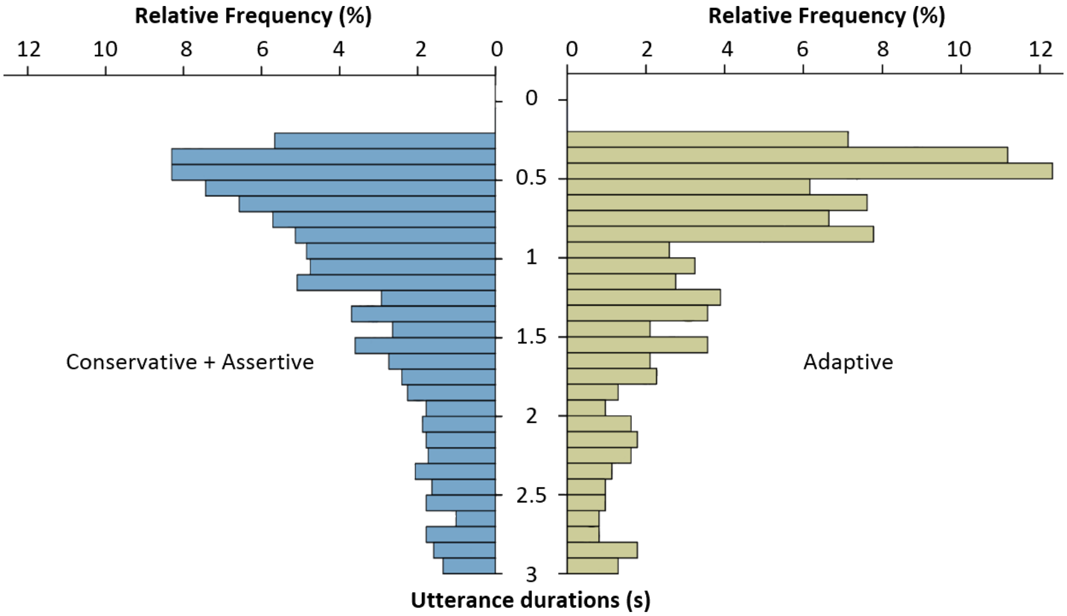


Figure 11: Utterance length distribution differences

The comparison shows the deviating distribution for the adaptive condition, in which utterances of approximately < 1 second appear to be more common than longer ones. From observations, awareness among participants that the robot seemed not hesitant to interrupt them, seemed to cause uncertainty with respect to the time participants had for their own turn. To increase the possibility for successful communication towards the other participant in case a cut-in is made by the robot, participants used seem to have used more concise utterances.

To gain insight to the opportunities for participants to speak, the number of average utterances per minute is compared between conditions. Only the conservative condition in contrast to the other two conditions shows a higher utterance count, *t*(38) = 2.34, *p* <.05, *r* =.36. This is made visable in Figure 12.

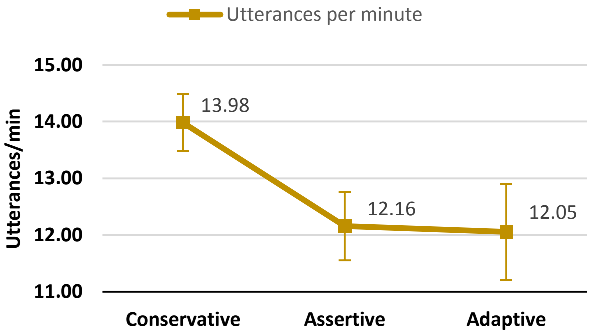


Figure 12: Utterances/min per condition

Gaze Cue

Timing

The data from the 85% successful turn take measurements in the conservative condition, is used to gain insight into the relative timing of the gaze cue and the silence cue. During or before the 1s silence after which the robot took turn, the point in time at which also the gaze cue is detected is calculated. The difference in time of occurrence between the two cues is referred to as delta-t. The data was filtered, and negative delta-t values < -1.60s were statistically considered outliers. A total of 130 delta-t measurements remained with a near normal distribution. The timing of a gaze cue (*M* = -0.53, *SE*= 0.05) is not significantly different from zero, t(129) = -0.98, *p* = *NS*, which means the cue occurred roughly at the same time as the start of a silence cue. In other words, 95% of gaze cue observations are equally likely to occur up to 1.24s (2*SD) before or after the start of a silence cue. This result indicates that there is no systematic order of appearance of the two cues. As can be observed from the distribution in Figure 13, delta-t has a relative wide range and can be positive as well as negative.

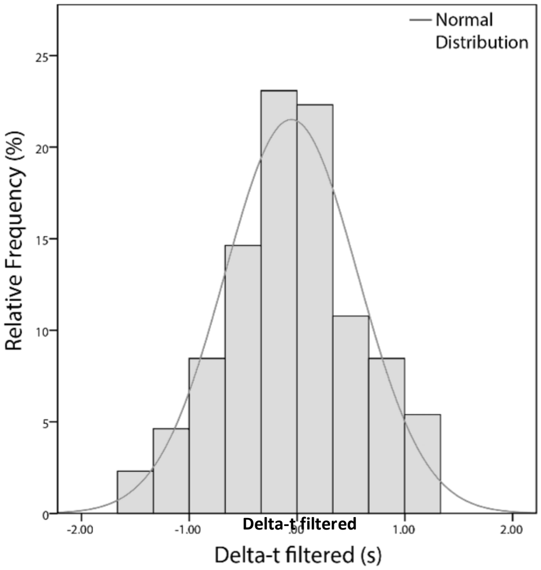


Figure 13: Filtered gaze cue timing

Adaptive Turn Take Model

For all turns taken by the robot, using the adaptive turn take model, 71% (*SD* = 5.3%) of them resulted in cut-ins. From the definition of the model's adaptive strategy, the distribution of applied rules as shown in Table 8 was extracted from the log files (the average distributions of all 7 individual sessions):

Table 8: Distribution applied adaptation rules

Adaptation rule	***Mean***	***St. Dev***	***Delay***
Pre-silence = FALSE & Cut-in = TRUE	67%	7%	Increase (+)
Pre-silence = FALSE & Cut-in = FALSE	24%	7%	Decrease (-)
Pre-silence = TRUE & Cut-in = FALSE	4%	2%	Increase (+)
Pre-silence = TRUE & Cut-in = TRUE	5%	3%	Decrease (-)

Noticeable is that 91% (*SD* = 14%) of all moments of gaze to the robot that led to a turn take, happened while the speaker was still talking (Pre-silence = FALSE). In addition to the filtered delta-t gaze time data, this finding suggests that a gaze cue is likely to be expected (just) before the silence cue. More than ⅔ of all turn takes in the adaptive condition, used an additive rule which increased the delay for subsequent turn takes. Despite the occurrence of a subtractive rule in almost ¼ of the remaining turns, a rapid increase in turn take delay is observed during the first half of the conversations in the adaptive condition. Having reached its maximum delay of 3 seconds, no more increase takes place during the second part of the conversations. With the exception of two experiment sessions that showed significant deviating trends from the total, the increase of delay over time is visualized and can be observed in Figure 14. The zoomed-in crop during the first 100 seconds of the experiment, shows the trend lines for several sessions and the loss of resolution from the moment the increments double in size.

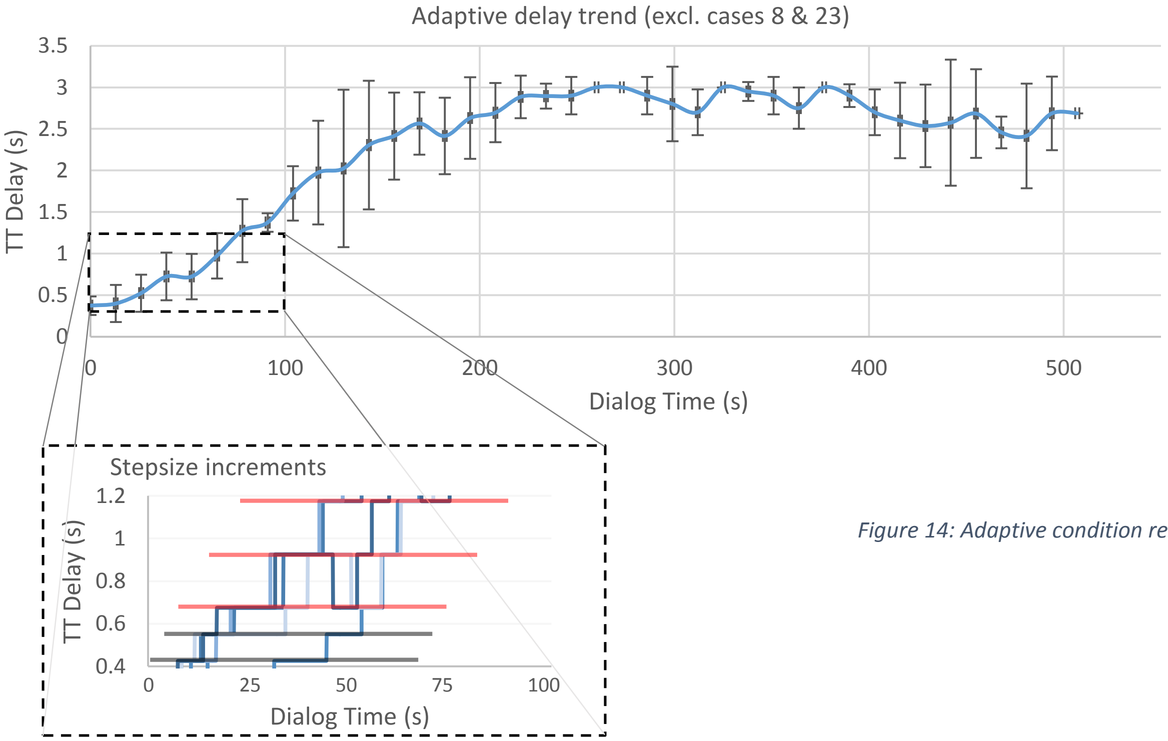


Figure 14: Adaptive condition results

It is suspected that a combination of the stair-wise increment and the large window range of 3 seconds, caused insufficient resolution to adapt in the turn take delay between 0.7 – 1.2s. These delay times partially fall within the typical range of turn take intervals measured in mutual conversations between participants as reported earlier.

The faction of cut-ins for turn take delay intervals can also be visualized. By grouping delays that are close to each other, a suitable interval (binsize) of 200ms is chosen to represent the cut-in fractions and respective standard errors. As can be seen in Figure 15, only for a few intervals the fraction of cut-ins was found to be below chance (< 0.5). In addition, some error bars reflect substantial uncertainties due to the lack of sufficient measurements within intervals. Noteworthy is the cut-in fraction minima of .42 for delays between 200 – 400ms (2nd bar), and maximums of > 0.9 that altogether lie within the range of typical turn intervals that were measured between human speakers.

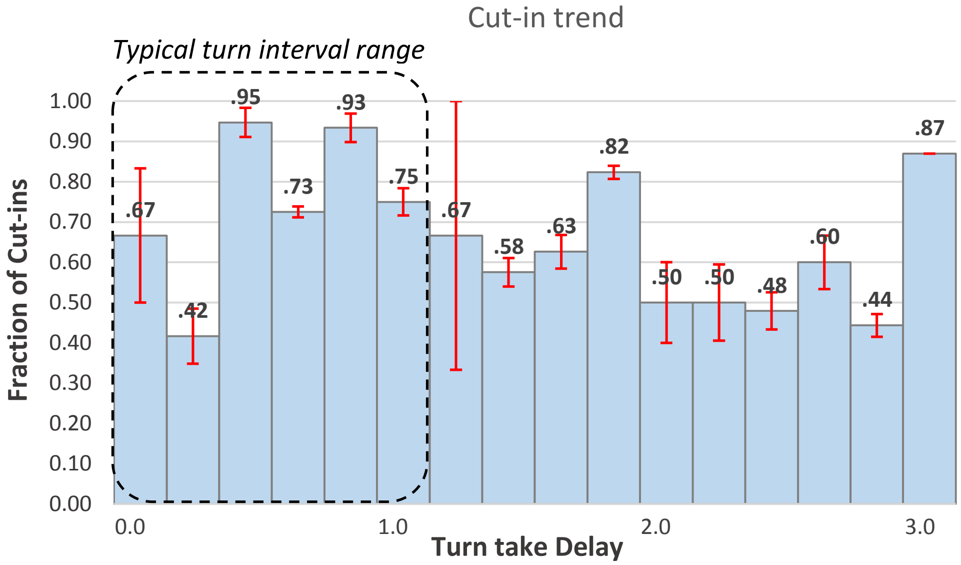


Figure 15: Delay duration versus cut-in risk

Analyses of Subjective Factors

To analyze the factors measured by the questionnaires, Cronbach's alpha reliability scores were calculated. Some individual items showed problematic correlation with other items of a factor, and were therefore excluded to raise the factor reliability score. Although lower values might be the result of a diversity of the constructs being measured within a factor, only factors with a score > 0.7 were retained due to the restricted sample size. From the SASSI questionnaire, only the habitability factor (α =.63) shows problematic reliability. For the Godspeed questionnaire however 3 factors; anthropomorphism (α =.67), animacy (α =.65), and perceived safety (α =.59), have questionable reliability scores. These measurements were therefore excluded from analyses, retaining eight from the original twelve measured factors. An overview of the reliability scores for all factors is given in Table 9.

Table 9: Factor reliability scores

Factor	***Number of retained items***	***Cronbach's Alpha (α)***
Response accuracy	9	.763
SASSI Likeability	9	.868
Cognitive demand	5	.734
Annoyance	3 (-2)	.811
Habitability [6]	4 (-1)	.625
Speed [6]	4	.780
Anthropomorphism	4 (-1)	.674
Animacy	3 (-1)	.649
Godspeed Likability	5	.847
Perceived Intelligence	5	.880
Perceived safety	2 (-1)	.594
Social Intelligence	4	.742

Analyses of variance

Comparing the means of the retained factors in an ANOVA, revealed a significant (2-tailed) effect of the three experimental turn take strategies on several measured factors. Differences in participant's judgment about the robot's response accuracy, $F(2,39) = 6.64$, $p <.01$, $\eta^2 = .50$, as well as speed were found, $F(2,22) = 8.94$, $p < .001$, $\eta^2 = .52$. The turn taking behavior also had a profound effect on the perceived intelligence of the robot, $F(2,36) = 3.82$, $p < .05$, $\eta^2 = .42$, and provoked annoyance, $F(2,39) = 9.18$, $p < .001$, $\eta^2 = .57$. Hence, also the likeability level from both the SASSI and Godspeed questionnaire differed between conditions, respectively $F(2,39) = 5.89$, $p < .01$, $\eta^2 = .48$, and $F(2,39) = 7.35$, $p < .01$, $\eta^2 = .52$. No effects ($p > .1$) were found for the factors cognitive demand, and social intelligence. With a mean level of 0.20 (SE = 0.13) on a scale from -3 to 3, the cognitive demand was neither high or low and not different from neutral (zero), $t(40) = 1.55$, $p = NS$. The robot's perceived social intelligence was with an average score of 3.67 (SE = 0.09) on a 5-point Likert scale only slightly more positive compared to neutral (three), $t(40) = 7.08$, $p <.001$, $r =.75$. Both factors reflect that the system can still be improved by increasing its social intelligence, while lowering the level of concentration that is required from people when interacting with Marvin.

[6] Dimensions were complemented with verified related item(s) from the REVU-NL questionnaire to increase dimension reliability (Dzikovska et al., 2011)

Planned contrasts

Two planned contrast analyses are used to gain insight into the differences between conditions. Table 10 sums the significant results of contrast I between the conservative and grouped assertive and adaptive condition, as well as the results of contrast II between the assertive and adaptive condition.

Response accuracy, perceived intelligence, and both SASSI and Godspeed likability were higher for the conservative condition than for either the assertive and adaptive conditions. Furthermore, annoyance was rated significantly lower in the conservative condition. These finding altogether point in the direction of preference for the conservative robot. Speed was however rated more positively for the adaptive and assertive conditions compared to the conservative condition.

The second contrast showed that Godspeed likeability, perceived intelligence, and speed were higher for the assertive condition compared to the adaptive condition.

Table 10: Contrast analyses results

Factors	*Contrast I (conservative vs. rest)*	*Contrast II (assertive vs. adaptive)*
Godspeed likability	*t*(39) = 3.4, *p* <.01, *r* =.48	*t*(39) = 1.75, *p* <.05 (1-tailed), *r* =.27
Perceived intelligence	*t*(36) = 2.2, *p* <.05, *r* =.36	*t*(36) = 1.71, *p* <.05 (1-tailed), *r* =.27
Speed	*t*(18.6) = -2.94, *p* <.01, *r* =.56	*t*(19.1) = 2.17, *p* <.05, r =.44
Annoyance	*t*(39) = -4.0, *p* <.001, *r* =.54	
Response accuracy	*t*(39) = 3.6, *p* <.001, *r* =.50	
SASSI likability	*t*(39) = 3.3, *p* <.01, *r* =.47	

A steep increase in annoyance between the conservative condition and the assertive and adaptive conditions becomes clear when plotting the data, as can be seen in Figure 16. Whereas both plots show a decline in likability, and the second plot also a decline in perceived intelligence for the assertive and adaptive conditions. The response accuracy is however perceived only marginally positive by participants in the conservative condition, and equally negative in assertive and adaptive condition.

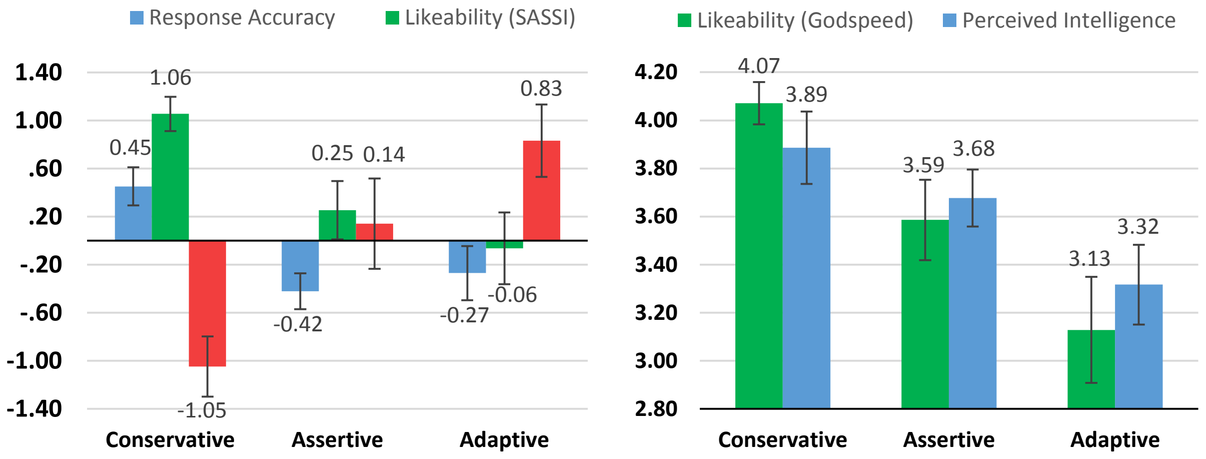


Figure 16: Significant factor differences per condition

In Figure 17 it can be observed that speed is the only variable that does not follow a linear trend downwards or upwards given the condition order from left to right (i.e. first conservative, secondly assertive, and thirdly adaptive). The order is based on the severity of turn take violations in which being slow is the mildest violation (conservative condition), and taking a turns too rapid (assertive condition) is less bad than bluntly cutting off someone (adaptive condition). Speed in the conservative condition is not different from neutral (zero), $t(12) = 2.09$, $p = NS$. The fact that the adaptive condition had non-consistent behavior due to variable turn take delays, caused in addition to many interruptions also more prolonged turn interval silences as the adaptive delay increased. Interruptions were thus alternated with long turn intervals. This alteration had a mediated effect on perceived speed of the system.

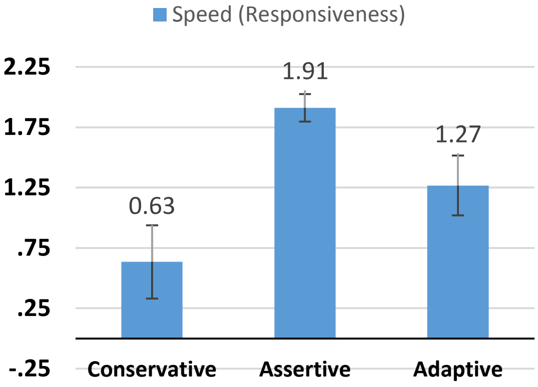


Figure 17: Differences in perceived system responsiveness

Interaction analyses

The age of participants, previous experiences with speech interfaces, the robot itself, or frequent gaming could influence the judgment and perception of the interaction with the robot. Several multivariate analyses have been performed on the significant factors from the ANOVA outcome. Multivariate analyses showed that age, gender, speech interface experience, and frequency of gaming had no effect on response accuracy, SASSI likability, annoyance, speed, Godspeed likability or perceived intelligence. The interaction analyses rule out any of the earlier mentioned variables as possible confounding factors with the experiment condition dependent variables. The only exception was the relation between robot experience and annoyance $F(1,32) = 4.322$, $p = .046$, $\omega = .119$, as shown in Figure 18.

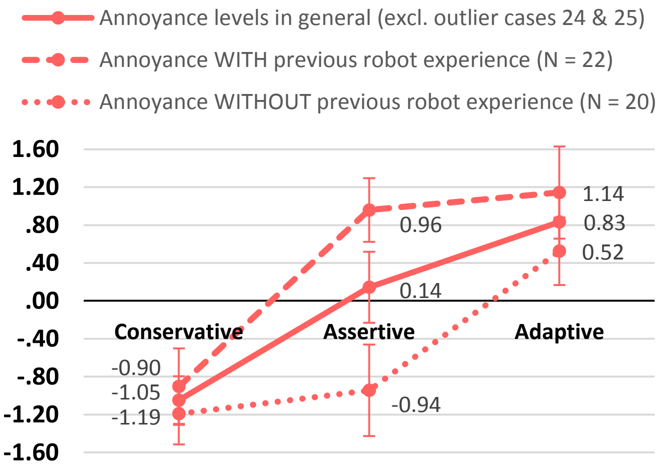


Figure 18: Effect of robot experience

When analyzing the effect of experiment conditions on annoyance levels, taking into account the variable ‘previous robot experience’ resulted in an increase of effect size; $F(2,35) = 6.661$, $p = .004$ $\omega = .276$ versus $F(2,32) = 7.853$, $p = .002$, $\omega = .329$. Since the ‘previous robot experience’ variable only significantly changes annoyance levels for participants in the assertive condition and not for participants in the other two conditions, there is no need to take into account the variable in subsequent analyses.

Multilevel analysis

As result of the experimental setting for having dialogs in participant pairs, the data from the questionnaire cannot just be assumed to represent data from independent samples. This is illustrated in Figure 19. A partial multilevel analysis is therefore done on several factors. No significant variance was found for intercepts originating from dyad grouping across participants, as show the test statistics for response accuracy, $var(u_{0j}) = 0.98$, $\chi^2(1) = 0.87$, $p = .351$, likability, $var(u_{0j}) = 0.09$, $\chi^2(1) = 0.313$, $p = .576$, and annoyance levels, $var(u_{0j}) = 0.36$, $\chi^2(1) = 1.66$, $p = .197$. Therefore it can be safely assumed that the data collected through questionnaires can be analyzed as regular data from independent participants.

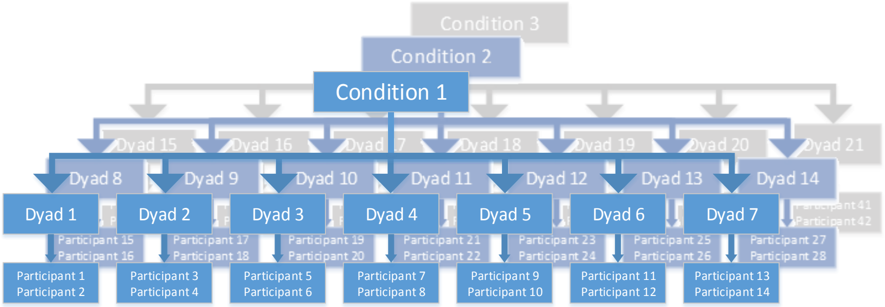


Figure 19: Multilevel model schematics

Individual item analyses

A total of eight items with semantic differences were added to the questionnaire. The non-continuous response scale of the individual items is not well suited for an ANOVA. The 5-scale response data of the items was therefore reduced into two nominal categories, which offered the possibility to do a Chi-square analysis. Neutral scores were split in half and added to both the nominal categories. This way each category indicated the score for the two opposites that was represented by the semantic difference. For the analysis Fisher's exact test was used, since the expected frequencies were close to five or even smaller. The test results showed that three of the original eight items showed differences in observed frequencies to expected frequencies, as result of the experiment conditions. The strongest effect was observed for participants' ratings on the robot's 'patience', $\chi^2 = 12.95$, $p < .001$, $Ø = .56$. In addition its 'politeness', $\chi^2 = 9.52$, $p < .05$, $Ø = .45$, as well as perceived 'dominancy' $\chi^2 = 6.58$, $p < .05$, $Ø = .41$, differed between conditions.

In order to gain an understanding of the differences of the three aforementioned items, non-parametric Mann-Whitney post-hoc analyses are performed on the original 5-scale response data. The analyses revealed that there are no significant differences between the assertive and adaptive condition. It is however the conservative condition that significantly differs from the assertive as well as the adaptive condition. Table 11 shows the central tendency scores (median) of the 5-scale (1 to 5) response data for each condition, in combination with the analyses results using Monte Carlo exact significance and Bonferroni correction ($\alpha = .025$):

Table 11: Individual item analyses

Items	***Medians***			***Comparison results***
	Conservative	***Assertive***	***Adaptive***	
Impolite – Polite	4.00	3.00	n/a	$U = 36.5$, $z = -3.01$, $p < .01$, $r = -.57$
Impolite – Polite	4.00	n/a	2.50	$U = 35.0$, $z = -3.07$, $p < .01$, $r = -.58$
Impatient – Calm	3.50	2.00	n/a	$U = 27.5$, $z = -3.33$, $p < .001$, $r = -.63$
Impatient – Calm	3.50	n/a	2.00	$U = 12.0$, $z = -4.05$, $p < .001$, $r = -.77$
Dominant – Reserved	3.00	1.50	n/a	$U = 36.0$, $z = -2.94$, $p < .01$, $r = -.55$
Dominant – Reserved	3.00	n/a	2.00	$U = 55.5$, $z = -2.08$, $p < .025$*, $r = -.39$

*(one-sided)

Analyses of the Open-ended Questions

Analysis of the responses to open questions of the questionnaire, were done using open coding followed by axial coding. This process consisted of conceptualizing (labelling) all individual responses with keywords and subsequently grouping similar labels to build up unique, more abstract response categories. Differences between the three experimental conditions have been analyzed by comparing the response frequencies of each category. An overview of the response frequencies can be found in APPENDIX B.

Expectations of the robot

Before the start of the experiment, participants indicated what they expected from the robot that was going to take part in a small group dialog. In the top five of most mentioned answers, responses reveal that participants fore mostly expected the robot to ask questions give relevant responses. Indirectly these statements, together with notions as '[It should] provide unique ideas and suggestions', imply that the robot is expected to listen and understand spoken language on a sufficient level. No specifics about the robot or the experiment were mentioned before participants finished the first part of the questionnaire. From the expectations, participants already anticipated the robot to act as a moderator and to look at speakers during conversation. This summarizes some of the main expectations of the dialog robot, which fortunately largely correspond to the behavioral design of robot for the current experiment.

General impression of the robot

The appearances of the robot were found to be aesthetically pleasing (e.g. "it looks cute"). In addition, the robot's voice was considered a bit robotic/artificial. From comments of several participants to the questionnaire, a preference was given for a voice like 'Siri' due to its human-like intelligibility, fluidity, and natural way of intonation that is perceived as considerate and characterized by politeness.

Properties like 'impatience' and 'not listening' do not occur for the conservative condition, but are mentioned equally for the other two conditions. Due to cut-ins and other negative judgements, one participant in the adaptive condition describes the interaction with Marvin as "*The conversation was forced, felt like he should have a whip in his hands while being the travel agent*". Peculiar is that 'helpful' is mentioned four times in the assertive condition compared to only once in the other conditions, which perhaps has to do with remarks about the robot increasing effective decision making due to its perceived fast and result-oriented attitude. Interesting is that two participants in the adaptive condition described the interaction as 'confusing'. This indicates that the variability in turn take timing for the adaptive condition was large enough to be perceived, causing confusion due to its seemingly unpredictable behavior. 'Machinelike' and 'synthetic voice' are terms only mentioned for condition. A noteworthy comment was given by one participant from the conservative condition stating that *"[Marvin was] Very helpful and giving useful information on destinations, so a good impression, but he doesn't perceive doubt very well*". This suggests that a higher level of intelligence was expected from the robot. The aforementioned comment in combination with other responses are remarkable since it seems that participants got more critical in the conservative condition (i.e. criticizing either advanced or trivial features).

Behavior of the robot

Participants' evaluation of the behavior of the robot is clearly different from participant's overall impression of the robot as reported earlier. The behavior between conditions is characterized distinctively. The conservative condition is favored significantly more (6x likable) to the other conditions (both scoring 1x likable). What completes the top three besides 'likeable' for the behavior of the robot in the conservative condition are intelligent (5x) and machinelike (3x). For the adaptive condition the top three consists of interrupting (5x), impatient (4x), and assertive (3x), while the assertive condition is defined as impolite (5x), too fast (4x), and dominant (3x). A promising quote of a participant from the conservative condition confirms the importance of responsive gaze behavior during multiparty dialog; "*His head movement was a strong cue and felt natural, when speaking it really looked like it paid attention*". As proof of successful manipulation, 'interrupting' is mentioned up to five times for the adaptive condition while it is mentioned only twice for the assertive turn take behavior. The perceived difference in reaction speed is only slightly different for the adaptive strategy versus the systematic fast turn take strategy in the assertive condition (3x vs. 4x too fast).

Adequate gazing behavior of the robot in the conservative condition, is supported by notions of 'looked at you' 3x, and 2x the notions of; liveliness, human-likeness, 'including both members', and 'natural head movements'. The positive behavior descriptives 'looked at you' and 'natural head movements' are only mentioned once, with in addition one negative notion stating 'nervous head movements'. This is probably due to shorter 'listening-state' periods of the robot in the adaptive and assertive conditions. Somewhat peculiar is the notion of 'unaccustomed' that is mentioned three times and only for the conservative condition. Instead of criticism, unaccustomed may refer to the unusual experience of having a conversation that involves a robot itself. Particular interesting is three times the notion of 'helpful' mentioned only by participants in the assertive condition. It presumably refers to the perceived advantage of fast turn taking, which increased the dialog effectiveness according to some participants.

Perceived personality of the robot

Just under 30% of the participants did not perceive any personality. From the remaining 70% of the participants, little over half did perceive a clear personality while the other lesser half vaguely perceived a personality. All participants however denoted human personality characteristics that described the robot's personality if it would possess one. Interesting to see is the notion of 'helpful' that is mentioned three times for both the conservative as well as the assertive condition, while omitted in the adaptive condition. From these responses it appears imperative that a verbal interactive system helpful to people, requires a turn take model that minimizes the chance of interrupting speakers during dialog. Unless there are well-considered deliberate reasons to allow for cut-ins, for instance when a robot should be able to intervene as moderator in a heated discussion or debate.

The attributes 'interested' and 'calm' are both mentioned twice in the top three of personality traits for the conservative condition. It is very likely that the gazing behavior of the robot combined with longer 'listening-states' periods underlie these notions. For the assertive condition one participant wrote "*I perceived some 'bossiness' like if he wanted to have control of the situation*". This comment refers to the perception of dominance which together with impatience and assertiveness resembled the top three of the assertive as well as the adaptive condition (respective frequencies 4x, 3x, and 2x). This indicates the thin line between the two conditions which was also reflected by the significant factors from the SASSI and Godspeed questionnaire. Peculiar is however the distribution for the stubbornness trait, which is mentioned twice for the conservative strategy, three times for the adaptive strategy, and not mentioned at all in the assertive strategy. Stubbornness most likely reflects erroneous choices made by the robot caused by experimenter control deficiencies. Rather amusing is the appearance of the term 'Salesman' for the adaptive condition, which is probably related to a combination of earlier mentioned personality characteristics typically associated with salesmen such as assertiveness, stubbornness, impatience (pushy), and dominance (decisive).

DISCUSSION

The aim of the study was to design a framework to evaluate the usefulness and effectiveness for a variety of turn taking strategies that use the gaze cue and differ in turn take delays. The gaze cue itself is also analyzed with respect to the absence of speech cue. Of interest as well was the influence of the robot's behavior, including its responsive gaze, on participants and the conversation itself.

Exponential distributions of participants' speech characteristics correspond with observations done by Jaffe and Feldstein (1970). Both researches likewise studied the duration of utterance and different types of silences between humans in dyadic conversation. Moments of silences between utterances from a single speaker, referred to as speaker silences, were found to maximally last 550ms (upper boundary 95% CI). Similarly turn intervals, the silence between the utterance of one speaker and the utterance of another speaker, were found to last up to 950ms (upper boundary 95% CI) with an average of 340ms. Given the evaluations from predominantly the conservative and assertive condition, it seems that there is not a particular sweet spot for turn take timing. Instead, from the human-to-human conversation measurements, a quite lenient turn take delay interval can be defined that satisfies general turn take effectiveness. Given that cut-ins occur on rare occasions, it is expected that people will not judge the system or the dialog too bad. Whenever a turn take by a system results in a cut-in after a noticeable moment of silence, it is likely that people are able link the system's 'mistake' to the occurrence of silence. Only serving as a guide line a delay of 550ms, based on the speaker silence upper boundary, can be used as a starting point to avoid cut-ins. Any variations may subsequently vary within a range from 340 – 950ms based on the participants' turn interval data.

Comparison of Turn Take Strategies

Conservative

It was hypothesized that turn take models with shorter delays and less speaker overlaps (cut-ins) would be more positively evaluated. Results show that the conservative turn take model which requires a 1s silence cue (i.e. 1 second delay before turn take) scores most favorable on aspects as response accuracy, perceived intelligence, and likeability. The conservative turn take behavior seemed adequate enough that it even caused participants to focus their attention more on other aspects of the robot, such as its motion fluency. High likability and low annoyance scores are crucial to the acceptance of similar turn taking strategies in human-robot verbal interaction. The response accuracy most accurately reflects the evaluation of the system's perceived naturalness. It evaluates how well the system does what people expect it to do in terms logical responses to situations. The conservative condition is the only condition that has a modest positive response accuracy score compared to the equally large but negative scores of the other conditions. Speed on the other hand is an important determinant of the efficiency of an interface. This is downside of the conservative condition. Its ratings of speed are neutral, with a trend to slightly positive, and show that the fixed delay of 1 second limits the dialog efficiency.

Assertive

Improper turn taking by speaking too soon or causing speech overlap, was expected to be tolerated by participants depending on the frequency and subjective severity, and with the exception for obvious interruptions. The assertive model, in which the robot took turn immediately after a short occurrence of both the silence cue and gaze cue, was designed to test the usability of the gaze cue for quick turn take strategies. Due to the additional requirement of gaze presence, the assertive strategy proved to be adequate enough to preserve a multiparty dialog. Nearly three times higher (positive) ratings of speed were found for the assertive condition. Due to the system's extraordinary fast turn taking however, ratings of response accuracy, SASSI factor likeability, and annoyance, were sub-optimal compared to the conservative condition. Participants described the interaction pace as goal-oriented, which increased the decision making effectiveness. Premature turn takes of a robot can apparently be perceived as a 'well intended' strategy of it to, for example, increase the interaction speed to force people in making decisions. The models' speed combined with mediocre but not severely negative evaluations of participants, provides evidence that turn take strategies with almost no perceivable delays can have a positive effect on the dialog.

Adaptive

Noticeable turn taking behavior was anticipated to influence also the dialog behavior of participants, as a result of people's ability to adapt and cope with new situations. Shorter utterance lengths and a slightly lower number of utterances per minute, suggest that participants actually changed their own speaking behavior as a result of the robot's behavior for particularly the adaptive condition. Participants most likely shortened and limited their utterances to enable concise and effective communication, while at the same time reducing the possibility for the robot to interrupt them while speaking. These results reflect that systematic mishaps in turn taking can influence the speaking behavior of other members of a dialog. Although several cut-ins were anticipated, the adaptive turn take model had a higher than expected cut-in percentage of 71%. The condition had the lowest rating for likeability and perceived intelligence, a mediocre speed rating, and the highest annoyance score. One participant in the adaptive condition also wrote; "*[I] realized how important silences are in a natural conversation*". Typical turn intervals between human speakers were studied by Heldner et al. (2011). Heldner and colleagues measured in total 6506 between-speaker intervals (turn-intervals), and found that the most frequent turn interval ranged up to 400ms. Based on their results it was expected that the adaptive turn take strategy would find a turn take delay in-between the delays of the two other conditions (i.e. 0s < adaptive < 1s).

For the adaptive condition, the cut-in distribution of Figure 15 depicts that the lowest cut-in rates were found for delays < 400ms, the highest cut-in rates (≥ 0.73) were found for the interval 400 – 1000ms, and a more or less slight decline in cut-in rate towards 0.5 was found for delays > 1000ms. Instead of representing the cut-in change at proper turn take moments, it is likely that the measured cut-in distribution represent the risk of speech overlap when turns are taken primarily at the wrong moments during conversation. Measurements of participants' speech showed that speaker silences typically didn't last longer than 550ms (upper boundary of the 95% CI). If any moment of gaze towards the robot was accompanied by speaker silence in the adaptive condition, its duration would typically not have lasted for longer than this upper boundary value. This presumably caused the relatively low cut-in fraction of .42 for turn take delays between 200 – 400ms. A similar fraction was therefore also expected for delays between 0 – 200ms, but could unfortunately not be confirmed due to insufficient measurements. It may however be clear that mid-sentence interruptions should definitely be avoided by any turn take strategy.

For typical turn take delays between 0 – 1s, a U-shaped cut-in probability distribution was actually expected. This was based on the intuitive sense that the requirement of a gaze cue in addition to the absence of speech for a certain duration, would facilitate successful turn taking of the robot and decrease the risk of cut-ins for the first half of the interval. Assuming that gaze cues reliably indicate turn take moments for the robot, participants were expected to have finished their turn. However, it was observed that after a certain time of speech absence, people tended to shortly speak in order to avoid awkward silences. This was done for example by using a 'filler' utterance (e.g. "sooo…"). Whether or not speech overlap with a filler utterance should be considered an actual cut-in, an increase in cut-in rates was again expected for delays in the second half of the typical turn take delay interval (near 1s). The cut-in risk for the remaining longer delays up to 3 seconds would subsequently depend on people's intention to actually re-take the floor again for a longer duration, to make an addition to their previous final turn after all. A cut-in rate approximating 50% was hence anticipated for these delay values, reflecting the neither high nor low chance of interrupting a participant given that turn-retake intentions are unknown. Even higher delays (≥ 3s) would eventually cause participants to take over the floor for sure, and therefore leads to a higher percentage of cut-in risk again for the robot.

The lack of additional rules to separate true gaze cues from ordinary moments of gaze, has very likely contributed to the unexpected turn take behavior and resulting cut-in rates of the adaptive strategy. The theoretic implementation-model of the adaptive strategy was built on the assumption that a gaze cue either precedes or lags behind a turn take interval (i.e. silence cue). Just like the other conditions, a moment of mutual gaze between a participant and the robot could trigger a turn take. Unlike the other conditions, no co-occurrence of absence of speech was needed. It was the combination of cues that implicitly separated moments of ordinary gaze from actual gaze cues in the conservative and adaptive condition. This caused the adaptive model to frequently take turn at improper moments during conversation. The combination with the possibility to take turn during ongoing speech, resulted in the repeated use of the same incremental-adaptation rule that led to a practically uninterrupted increase of the adaptive turn take delay (i.e. unstable loop). An adaptive strategy that only considers moment of gaze as turn take cue when joined with a moment of speaker silence, should not be affected by the same issue. In addition, the variable delay could be adjusted by a 'silenced' cut-in feedback loop. This way, the adaptive turn take strategy checks if the current delay would result in a cut-in, but prevents a turn in the case this is true in order to avoid mid-sentence interruptions.

Gaze

Usefulness of the gaze cue

The requirement of a gaze cue to take turn made sure that both the conservative and assertive condition functioned properly, and resulted in clear perceptional differences for the robot's behavior. Turn take strategies that used the gaze cue to determine turn taking moments, were hypothesized to increase turn taking reliability, but not decrease turn take delays due to the necessary absence of speech period. The inclusion of gaze did reduce in particular the chance of cut-ins for situations where people had a mutual conversation first, thereby temporarily excluding the robot from the conversation. For the conservative and assertive models, the difference between plain gaze and a turn take gaze cue was defined by the co-occurring absence of speech at the moment of gaze. Timing measurements reflect that the gaze cue occurs near simultaneous with the start of the silence cue. However, large deviations in timing show that the gaze cue is equally likely to occur just before as well as after a participant is finished talking. This is not in line with the expectation that the gaze cue would systematically occur just after the start of the silence cue. The hypothesis was based on the intuitive sense that participants would prefer looking back at another human participant first at the end of a turn, to make sure if he/she wants to say anything else before the robot might start speaking. A possible explanation for the measured gaze behavior, is that a participant at the end of a turn already knows if the other participant has any intention to take over the turn (by means of non-verbal cues). If this is not the case, the speaking participant already looks at the robot just before finishing speaking. Instead of being an early predictor itself and speed up a turn take model, gaze in combination with the silence cue mainly enables more reliable detection of turn take moments. An overview of achievable cut-in rates for fixed silence cue durations (thresholds) without any additional cues are found in the End-of-Turn accuracy figure in the study of Raux and Eskenazi (2012). Due to the increased reliability by taking into account gaze cues, silence cues that are normally needed to achieve certain cut-in rates, may prove to be shorter. No proof for this is however presented in the current study.

Responsive gaze

The gazing behavior of the robot was aimed and expected to make participants feel personally addressed, when the floor was released after a question or remark from the system. From responses it appears that the head turning behavior of the robot was perceived to be a strong and natural cue. It created the illusion that the robot pays attention, listens to speakers by looking at them, and by alternating its gaze made sure to include both members in the conversation. Hence it can indeed be assumed that responsive gaze increased the likability and perceived liveliness. Reciprocal gaze, if properly implemented, seems to be valuable behavior that provides people with implicit feedback about the system's state and generally improves the interaction experience. Although the gazing behavior was deliberately programmed not to look at a participant when the robot introduced a new conversation topic, its visual attention by means of a head turn did directed itself to one of the participants during follow-up inquiries. The first participant to respond in an effort to answer the question that was asked, was practically always the person that the system had released the floor to by looking at them. This is also consistent with observations by Bohus and Horvitz (2010), referred to as first-respondent behavior after turn yields of their system. In rare cases it happened that the robot looked away from a talking participant due to non-speech related noise that was picked up by the microphone of the other participant. Rather striking was that this caused the speaking participant to stop talking, even though the robot stayed silent. The participants also tended to look at Marvin (the robot) first, and waited for Marvin to gaze back at them before addressing it. It appears that people are extremely sensitive to this cue, even when responsive gaze comes from an artificial system.

Constraints

Experiment setup

A slightly underestimated problem during the experiments, were moments of the poor intelligibility of the robot due to shortcomings in the speech synthesizer. Whenever participants had trouble to understand the robot, despite instructions they instinctively asked it to repeat its last utterance. This was however not possible for the current implementation, and caused some moments of confusion during several conversations.

15% – 17% of the measurements in the conservative and assertive condition were considered outliers due to deviating turn take delays. These delays lasted longer than the predetermined maximum duration for each condition (respectively 1s and ≈ 0s). This was caused by gaze cue detection failures before or during the time of the silence cue, as result of problematic face recognition and head-pose estimation. Analyses of the log files revealed that on average 71% (SD = 10%) of the time a face was successfully detected from camera frames that got processed during the experiment while the robot's head was not moving. Ideally this number would be closer to one hundred percent given the fixed positioning of participants in the current experiment setup. Recognition errors were mainly caused by participants' own head motions, certain hairstyles, and clothing/wearables that introduced patterns that 'confused' the face recognition. The combination of a short and tall participant in a dialog session also caused problematic face detection, and additionally resulted in perceivable lag for the robot's responsive gaze behavior. This was due to the time needed to relocate a face in the camera stream after a head turn, by subsequently adjusting the robot's gaze either upwards or downwards to properly look at the participant.

Turn take models

Generalizability

To realize a multi-person setting, a discussion group of 3 members that included the robot was used. The generalizability of the current turn take models are limited. To increase the actual employability for a dialog robot, requires support for open-world interaction. Aside from necessary adjustments for technical feasibility (fast multi-face-recognition, -tracking, and -head-pose estimation) the main challenge is to detect, track and keeping apart the aural and visual input of multiple people. The current turn take strategies also lack behavioral strategies to recognize and deal with multiple observers, and different attention dynamics by means of gaze cues (e.g. joint attention towards areas of interest). In addition, different conversation strategies are likely needed to support various group sizes and formation changes of interlocutors due to unrestricted mobility.

Implementation remarks

None of the turn take models took into account the sequence of the dialog course; whether a participant responded directly to the robot by staying focused on it and excluding the other participant, or when a participant included the other group member first before gazing back at the robot to return an utterance. Because any instance of gaze in itself was enough to trigger a turn take after some variable delay in the adaptive condition, it was only this condition that showed unexpected turn taking behavior. In contrast to the adaptive condition, the assertive condition did have the intended effect. Nevertheless it would have been beneficial if a slightly different turn take implementation was used. Due to uncertainty of the interrelationship between the absence of speech cue and the gaze cue, the implemented turn take window allowed for some delay between the gaze and the silence cue. Instead of the already simplified window in which the order of cues did not matter, an even simpler design without any interval window at all could have been used. The reason for this is that the current definition of the assertive model (window) is asymmetric. A silence before gaze scenario will always result in a turn take at which both cues are valid. In the contrary, a gaze before silence scenario allows the gaze cue to not be valid anymore at the moment of turn take. Hence, a more informative and consistent implementation can be achieved by requiring the cues to be valid at the same time for a brief moment, regardless of their duration and order of appearance. This implementation can provide even more data to supports findings for the simultaneous occurrence and usability of the two cues in turn taking.

It became evident from the feedback received from participants, that there was a dichotomous view on the robot's behavior as result from its turn take strategy. While participants in the conservative condition characterized the behavior of the robot as a 'consultant willing to listen', the analogy was made with an 'eager salesman/rigorous moderator' for the assertive and adaptive condition. This was due to increased levels of dominance attributed to the robot's personality. Besides shorter turn take timings, previous work by Hung, Huang, Friedland, and Gatica-Perez (2008) showed that speaking length strongly correlates with dominance. Differences in speaking time for the robot between experimental conditions may have influenced the dominance ratings. With a standard deviation of 4% in all conditions, an average of 28% of the total dialog time was used by the robot to speak in the conservative condition. This is close to the ideal and equal speaking-time distribution of ⅓ of the total time for a group of 3 members. This is however quite a contrast to the mean of 50% and 54% of the time when the floor was occupied by the robot in respectively the adaptive and assertive conditions. The percentages in the latter two conditions are higher since participants were frequently forced to stop speaking prematurely, and deprived of the chance to finish their turn.

Recommendations

Turn take robustness

One way to improve a turn take model is to increase its speed by taking into account what kind of response is to be expected. A simple but illustrative comparison is the difference in answers that is to be expected from people after open questions or closed questions such as confirmations. Since it is very likely that a single short utterance follows a closed question, the chance of a cut-in for a system at re-taking the turn is low. This is in contrast with answers to open questions that are more likely to contain hesitations. Without taking more risk, models with variable turn take latencies can therefore alternate in these situations between a more assertive or rather conservative turn taking strategy. With a conservative approach being suitable for open ended questions and an assertive strategy being suitable for closed ended questions, this is also mentioned by Raux and Eskenazi (2012). They argue that this kind of implementation leads to faster responses whenever possible, and replicates a behavioral pattern that also exists between human interlocutors.

Furthermore, all turn take strategies depend on the fundamental ability to detect turn take opportunities. Robust detection of so called floor releases is the first step that determines the quality of a model. Besides the absence of speech as a cue in combination with the gaze behavior of the current speaker, being able to track the gaze of multiple members in a dialog group is expected to further increase turn take robustness. This is comparable to the ability of humans to scan and infer the gaze direction of other individuals during conversation using our peripheral view.

A different approach to increase turn taking reliability is reported by Bohus and Horvitz (2011b). They suggest to infer information about the interaction context, using easy available knowledge such as ‘who was the last speaker’ or ‘what is the current visual focus of the system’. Taking into account this higher order dialog context can significantly increase accurate detection of floor releases. Almost perfect detection is still a long way to go, and is due to the unpredictable behavior of people perhaps not even a realistic goal. Hence, any dialog system should have a mechanism to cope with unfortunate turn takes. The most obvious solution in these situations for a robot is to interrupt itself, and stop talking to give the human conversation member the chance to retake the turn and properly finish it. Failing to embed this kind of behavior leads to annoyance as was observed in the current study. Depending on the total number and length of utterances (chunk size) that the system had intended to say, a proper ‘recovery’ process needs to be defined to smoothly go back to a regular system state. Wilcock (2012) proposes for example a model that saves several progression parameters, such as pre-interruption system state, and utterances that remain to be verbalized. In addition, the system explicitly acknowledges the erroneous turn take by making a quick and brief apology towards the ‘rightful’ floor holder, after which it stays silent to await further input. This is likely to improve the likability and acceptance of any dialog system.

Non-verbal behavior

A different type of extension specifically applies to the non-verbal behavior of the robot. Richer non-verbal feedback promises more sophisticated interaction between people and artificial systems. Based on the suggestion of a few participants, it could be useful that a robot can somehow indicate when it is about to say something (i.e. ‘intention to speak’ parameter). For the current experiment no such feedback was given, nor was feedback provided about the system’s understanding of the verbal communication, since the robot only ‘pretended’ to listen (Wizard of Oz). This led some participants to feel uncertain about the level of understanding of the robot as they addressed it. Advisable for fully autonomous dialog systems is therefore the implementation of for example subtle head-nods, non-distractive changing eye-colors, or similar facial characteristics that can be used to represent the artificial equivalent of back-channeling. More specifically, the feedback could represent real time variations in specific variables like the audibility level of a speaker, speech recognition success rate, or the probability of correct semantic understanding. The options to complement the verbal speech abilities of a robot, are not limited to only small behaviors. Clearly perceivable expressive behavior such as gestures can also be added to convey certain intentions during dialog.

Adding gestures a complex challenge, due to the risk of being a distractor or break down the effect of other robot behavior aspects. Being aware of the potential negative influences of gestures, led to the exclusion of body movements for the robot in the current study. Nevertheless it remains interesting for future studies to investigate in what way gestures could affect verbal interaction between humans and a robot in addition to gaze cues. Adding motions to improve people’s perception of the anthropomorphism and liveliness of a robot proposes however many challenges. Successful implementation depends among other factors on simple limitations such as motion fluency, speed, degrees of freedom, as well as the delicate timing and synchrony with other robot behaviors. Beat gestures for example are hand and arm movements that can accompany utterances. A proper implementation requires however to take into account the accurate synchronization with pitch accents in the utterances pronounced by the robot (Wilcock, 2012).

Dialog capabilities

To prevent miscommunications that would degrade the interaction experience, a Wizard of Oz method with a pre-scripted dialog course was used. For real-world systems, semantic processing will however be necessary. The lack of derived semantics and pragmatics is a clear weakness for any general applicable dialog system as pointed out by Trafton et al. (2008). In practice however, many errors already occur due to inadequate performance on lower processing levels such as speech recognition modules. In the current experiment the robot for example asked participants to elaborate on some utterance. Participants verbalizing indecisive thoughts like “Uuhm well…I guess that…maybe…”, led the robot to a state in which meaningful speech was supposedly detected and a turn take upon silence was justified. A follow up question like “why do you think that?” therefore made no sense since no actual thoughts had been shared yet. Language processing would for example be useful to differentiate between meaningful and meaningless utterances. Processing the information from a conversation between people should thereby also enable a dialog system to estimate if there is difficulty reaching consensus between members of a discussion group for example.

Evaluating turn taking

The evaluations per condition show that more optimal dialog experiences are related to avoidance of cut-ins and acceptable turn take delays. To indicate various degrees of improper turn taking, the severity of cut-ins is hard to define. Clearly overruling and interrupting a speaking person versus taking over the turn whenever a speaker briefly stops talking, illustrates a clear perceivable distinction for cut-in severeness. It can therefore be said that the assertive condition predominantly made cut-ins, improperly taking turn during a brief moment of absence of speech. In comparison, the adaptive condition more often caused barge-ins due to the robot’s ability to take turn despite on-going speech. These type of interruptions resulted in more severe negative evaluations. Like cut-ins and barge-ins, uncoordinated hand gestures while talking, or looking away from an addressees, are role violations with regard to conversation disciplines (i.e. guidelines). So far these classifications rely on subjective evaluations only. For future research it could however be useful to devise a violation scale as a way to more objectively define how well a dialog system scores on conversation disciplines.

CONCLUSION

In the current study a framework was designed and successfully implemented to study turn taking in multiparty settings with a humanoid robot. Taking into account the gaze behavior of people during conversation in addition to their speech behavior, proved to be a useful cue combination to enable reliable turn taking in a group with two human interlocutors.

Three different gaze-based turn take strategies that mainly differed in their requirements regarding the absence of speech cue duration, were evaluated for their usability and effectiveness. An assertive turn take model that only needed a brief moment of silence to take turn, was implemented to test if the additional requirement of a gaze cue enabled a fast turn take strategy to manage turn taking successfully. A multiparty conversation was indeed possible, however the increased turn take speed frequently caused the robot to take turn too soon. As expected, this led to the interruption of turns from participants and resulted in less favorable ratings of the robot's behavior.

A conservative turn take strategy that required a relatively long absence of speech cue, was the most favored strategy. This was due to the low number of wrongful turn takes, and the perception that the robot took the time to 'listen'. The latter was also caused by the gaze behavior of the robot itself, which was a compelling cue for participants. Responsive gaze was engaging and had a positive influence on participants as they felt personally addressed when the robot turned its head and asked questions.

An adaptive turn take model that was designed to automatically adjust its speed depending on its turn take proficiency, resulted in unexpected behavior. Not waiting for a period of absence of speech, the model frequently interrupted ongoing speech of participants. This condition was least favored and led to high levels of annoyance and lower levels of perceived intelligence. Together with the assertive condition, personality traits like dominancy and impatience were observed by participants. The adaptive turn take model as well as the assertive model showed that a robot's perceived personality can be significantly influenced by changes to a single parameter like turn take timing.

The conservative strategy also provided relative timing measurements of the participant's gaze cue in respect to the absence of speech cue. Analysis showed that the gaze cue can be expected to occur roughly at the same time as the start of the silence cue. However a large deviation in observations revealed that the gaze cue is equally likely to precede or fall behind the start of the silence cue. Gaze as turn take cue can therefore increase the performance of turn take strategies in combination with other cues, but will not necessarily led to faster turn taking. Although quick turn taking can contribute to a better interaction experience, evaluations in the current study suggest that increasing turn taking robustness is more important and can be achieved by gaze-based turn taking strategies.

REFERENCES

Bartneck, C., Kulić, D., Croft, E., & 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.

Bennewitz, M., Faber, F., Joho, D., Schreiber, M., & Behnke, S. (2005a). *Integrating vision and speech for conversations with multiple persons.* Paper presented at the Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on.

Bohus, D., & Horvitz, E. (2009a). *Learning to predict engagement with a spoken dialog system in open-world settings.* Paper presented at the Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.

Bohus, D., & Horvitz, E. (2009b). *Models for multiparty engagement in open-world dialog.* Paper presented at the Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.

Bohus, D., & Horvitz, E. (2010). *Facilitating multiparty dialog with gaze, gesture, and speech.* Paper presented at the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction.

Bohus, D., & Horvitz, E. (2011a). *Decisions about turns in multiparty conversation: from perception to action.* Paper presented at the Proceedings of the 13th international conference on multimodal interfaces.

Bohus, D., & Horvitz, E. (2011b). *Multiparty turn taking in situated dialog: Study, lessons, and directions.* Paper presented at the Proceedings of the SIGDIAL 2011 Conference.

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems, 42*(3), 167-175.

Cassell, J. T., O; Prevost, Scott. (1998). Turn taking vs. Discourse Structure: How Best to Model Multimodal Conversation. *Machine Conversations*, 143-154.

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. *Handbook of mixed methods in social and behavioral research*, 209-240.

Csapo, A., Gilmartin, E., Grizou, J., Han, J., Meena, R., Anastasiou, D., . . . Wilcock, G. (2012). *Speech, gaze and gesturing: multimodal conversational interaction with Nao robot.* Paper presented at the ENTERFACE12 Summer Workshop-final report.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology, 23*(2), 283.

Dzikovska, M. O., Moore, J. D., Steinhauser, N., & Campbell, G. (2011). *Exploring user satisfaction in a tutorial dialogue system.* Paper presented at the Proceedings of the SIGDIAL 2011 Conference.

Goffman, E. (1955). On face-work: An analysis of ritual elements in social interaction. *Psychiatry, 18*(3), 213-231.

Hassenzahl, M., Platz, A., Burmester, M., & Lehner, K. (2000). *Hedonic and ergonomic quality aspects determine a software's appeal.* Paper presented at the Proceedings of the SIGCHI conference on Human Factors in Computing Systems.

Heldner, M., Edlund, J., Hjalmarsson, A., & Laskowski, K. (2011). *Very Short Utterances and Timing in Turn-Taking.* Paper presented at the INTERSPEECH.

Hone, K. S., & Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering, 6*(3&4), 287-303.

Hung, H., Huang, Y., Friedland, G., & Gatica-Perez, D. (2008). *Estimating the dominant person in multiparty conversations using speaker diarization strategies.* Paper presented at the Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on.

Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue* (8). Academic Press.

Johansson, M., Skantze, G., & Gustafson, J. (2013). Head pose patterns in multiparty human-robot team-building interactions. *Social Robotics*, 351-360.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22-63.

Klotz, D., Wienke, J., Peltason, J., Wrede, B., Wrede, S., Khalidov, V., & Odobez, J.-M. (2011). *Engagement-based multiparty dialog with a humanoid robot.* Paper presented at the Proceedings of the SIGDIAL 2011 Conference.

Matsusaka, Y., Fujie, S., & Kobayashi, T. (2001). *Modeling of conversational strategy for the robot participating in the group conversation.* Paper presented at the INTERSPEECH.

Michalowski, M. P., Sabanovic, S., & Simmons, R. (2006). *A spatial model of engagement for a social robot.* Paper presented at the Advanced Motion Control, 2006. 9th IEEE International Workshop on.

Mileounis, A., Cuijpers, R., Barakove, E. I. (2015). C*reating Robots with Personality: The Effect of Personality on Social Intelligence*. Artificial Computation in Biology and Medicine, 9107, 119-132

Mutlu, B., Forlizzi, J., & Hodgins, J. (2006). *A storytelling robot: Modeling and evaluation of human-like gaze behavior.* Paper presented at the Humanoid Robots, 2006 6th IEEE-RAS International Conference on.

Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., & Hagita, N. (2009). *Footing in human-robot conversations: how robots might shape participant roles using gaze cues.* Paper presented at the Proceedings of the 4th ACM/IEEE international conference on Human robot interaction.

Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2009). *Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior.* Paper presented at the Proceedings of the 4th ACM/IEEE international conference on Human robot interaction.

Raubenheimer, J. (2004). An item selection procedure to maximize scale reliability and validity. *SA Journal of Industrial Psychology, 30*(4), p. 59-64.

Raux, A., & Eskenazi, M. (2009). *A finite-state turn-taking model for spoken dialog systems.* Paper presented at the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Raux, A., & Eskenazi, M. (2012). Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing (TSLP), 9*(1), 1.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, 696-735.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk, 71*, 93.

Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2004). *Where to look: a study of human-robot engagement.* Paper presented at the Proceedings of the 9th international conference on Intelligent user interfaces.

Tannen, D. (2012). Conversational signals and devices. *A Cultural Approach to Interpersonal Communication: Essential Readings*, 157.

Trafton, J. G., Bugajska, M. D., Fransen, B. R., & Ratwani, R. M. (2008). *Integrating vision and audition within a cognitive architecture to track conversations.* Paper presented at the Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction.

Turunen, M., Hakulinen, J., Melto, A., Heimonen, T., Laivo, T., & Hella, J. (2009). *SUXES-user experience evaluation method for spoken and multimodal interaction.* Paper presented at the INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009.

Ward, N. G., Fuentes, O., & Vega, A. (2010). *Dialog prediction for a general model of turn-taking.* Paper presented at the INTERSPEECH.

Wilcock, G. (2012). *WikiTalk: A spoken Wikipedia-based open-domain knowledge access system.* Paper presented at the 24th International Conference on Computational Linguistics.

Yoshikawa, Y., Shinozawa, K., Ishiguro, H., Hagita, N., & Miyamoto, T. (2006). *Responsive Robot Gaze to Interaction Partner.* Paper presented at the Robotics: Science and systems.

APPENDIX A: Questionnaire constructs and example

Based on SASSI survey:

Response accuracy

1. Marvin is accurate
2. Marvin is unreliable
3. The interaction with Marvin is unpredictable
4. Marvin didn't always do what I wanted
5. Marvin didn't always do what I expected
6. Marvin is dependable (reliable)
7. Marvin makes few errors
8. The interaction with Marvin is consistent
9. The interaction with Marvin is efficient

Likeability (affect)

10. Marvin is useful
11. Marvin is pleasant
12. Marvin is friendly
13. I was able to recover easily from errors
14. I enjoyed using Marvin
15. It is clear how to speak to Marvin
16. It is easy to learn to use Marvin
17. I would use this system
18. I felt in control of the interaction with Marvin

Cognitive demand

19. I felt confident using Marvin
20. I felt tense using Marvin
21. I felt calm using Marvin
22. A high level of concentration is required when using Marvin
23. Marvin is easy to use

Annoyance

24. The interaction with Marvin is repetitive
25. The interaction with Marvin is boring
26. The interaction with Marvin is irritating
27. The interaction with Marvin is frustrating
28. Marvin is stubborn

Habitability (transparency)

29. I sometimes wondered if I was using the right word
30. I always knew what to say to Marvin
31. I was not always sure what Marvin was doing
32. It is easy to lose track of where you are in an interaction with Marvin
33. I knew what I could say or do at each point in the conversation with Marvin

Speed

34. The interaction with Marvin is fast
35. Marvin responds too slowly
36. It took Marvin too long to respond to my statements
37. Marvin responded quickly

Based on Godspeed survey:

Anthropomorphism

38. Fake - Natural
39. Machinelike - Humanlike
40. Unconscious - Conscious
41. Artificial - Lifelike
42. Moving rigidly - Moving elegantly

Animacy

43. Dead - Alive
44. Stagnant - Lively
45. Mechanical - Organic
46. Inert - Interactive
47. Apathetic - Responsive

Likability

48. Dislike - Like
49. Unfriendly - Friendly
50. Unkind - Kind
51. Unpleasant - Pleasant
52. Awful - Nice

Perceived Intelligence

53. Incompetent - Competent
54. Ignorant - Knowledgeable
55. Irresponsible - Responsible
56. Unintelligent - Intelligent
57. Irrational - Rational

Perceived safety

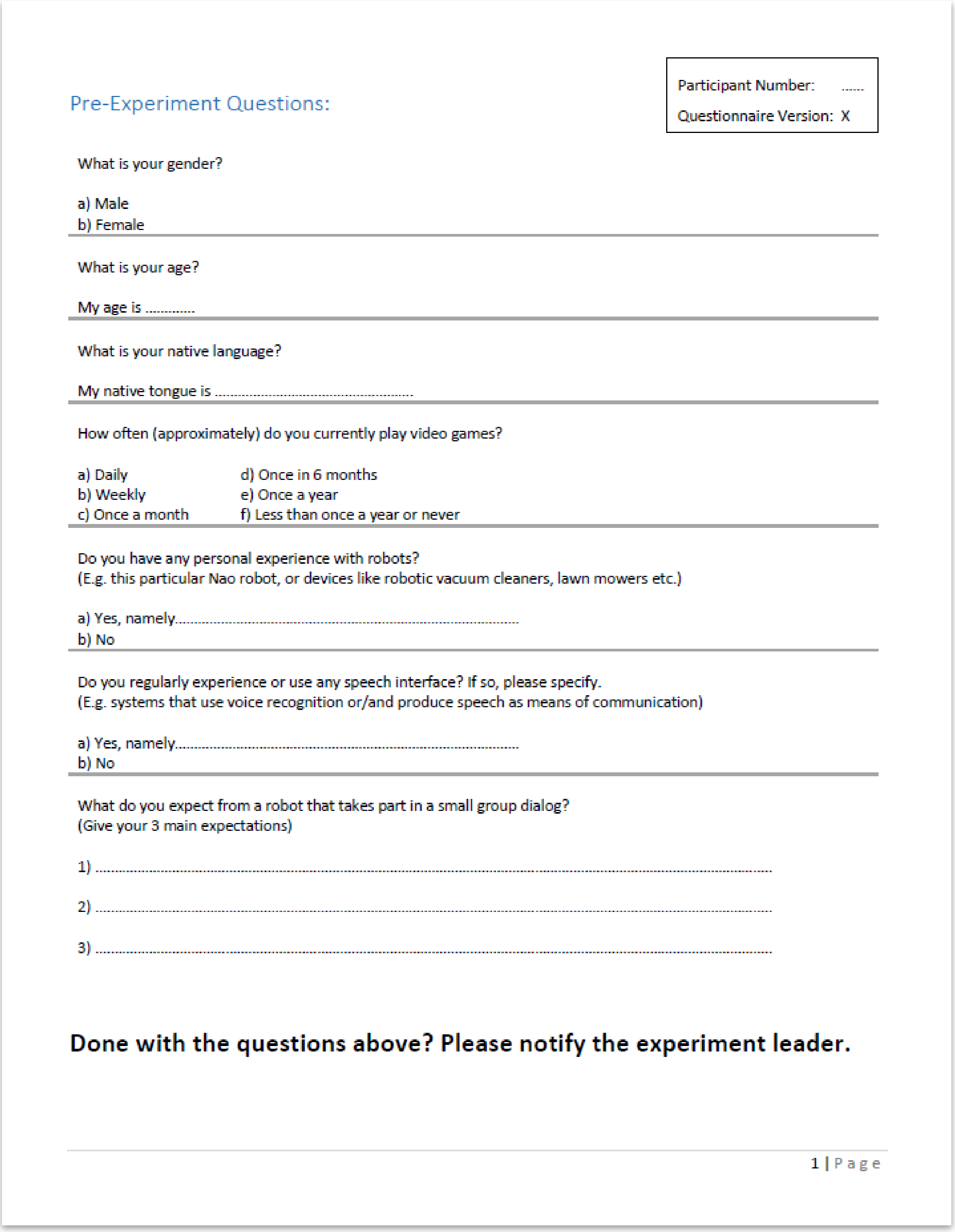
58. Anxious - Relaxed
59. Agitated - Calm
60. Unimpressed - Surprised

Social Intelligence

61. Uncooperative - Cooperative
62. Situation unaware - Situation aware
63. Unsupportive - Supportive
64. Discourageable - Persuasive

Isolated items

65. Uninterested - Curious
66. Impolite - Polite
67. Assertive - Conservative
68. Distracted - Attentive
69. Impatient - Calm
70. Dominant - Reserved
71. Extrovert - Introvert
72. Passive - Active



Participant Number: ......

Questionnaire Version: X

Pre-Experiment Questions:

What is your gender?

a) Male
b) Female

What is your age?

My age is ....

What is your native language?

My native tongue is ....

How often (approximately) do you currently play video games?

a) Daily
b) Weekly
c) Once a month
d) Once in 6 months
e) Once a year
f) Less than once a year or never

Do you have any personal experience with robots?
(E.g. this particular Nao robot, or devices like robotic vacuum cleaners, lawn mowers etc.)

a) Yes, namely..
b) No

Do you regularly experience or use any speech interface? If so, please specify.
(E.g. systems that use voice recognition or/and produce speech as means of communication)

a) Yes, namely..
b) No

What do you expect from a robot that takes part in a small group dialog?
(Give your 3 main expectations)

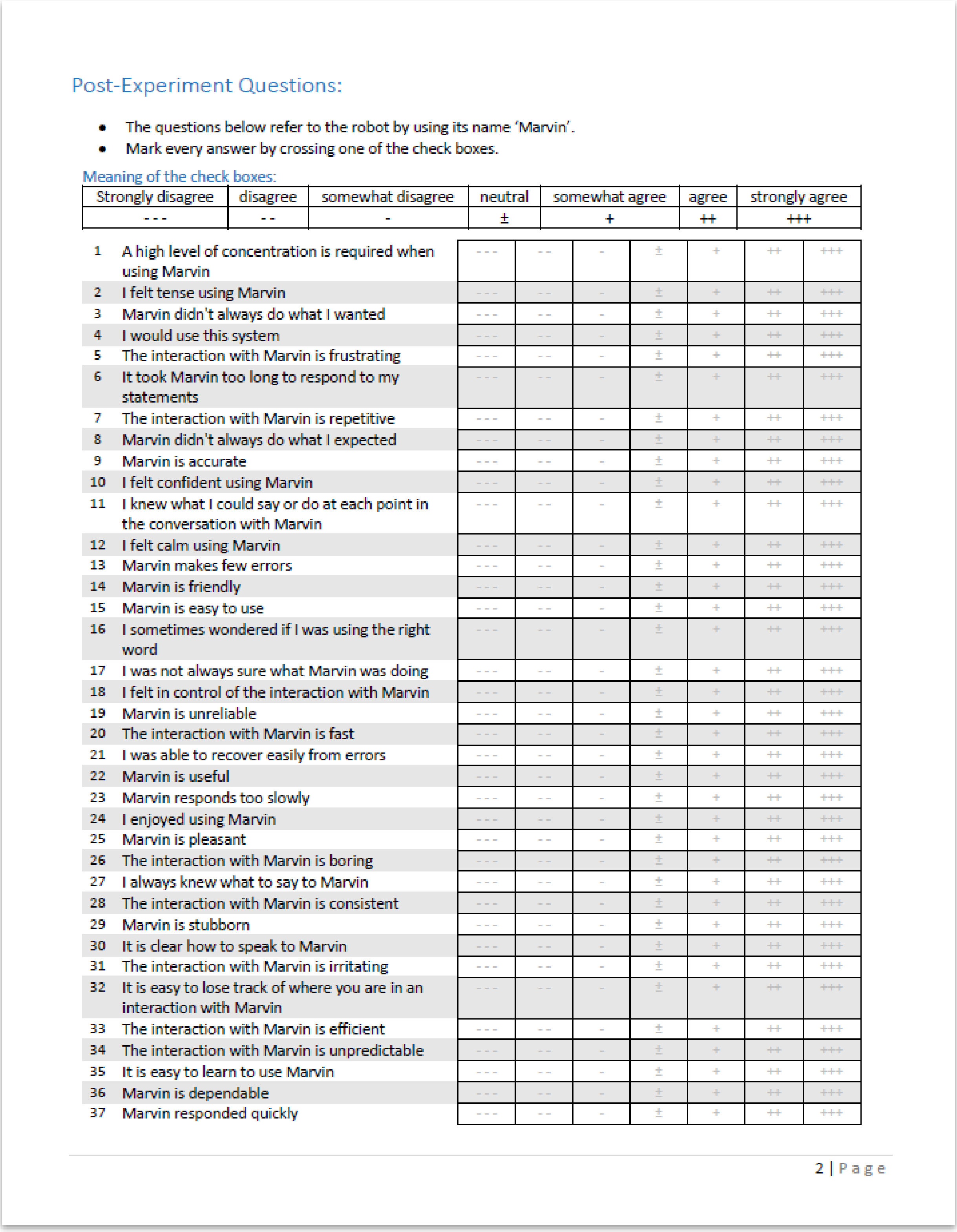
1) ....

2) ....

3) ....

Done with the questions above? Please notify the experiment leader.

1 | Page



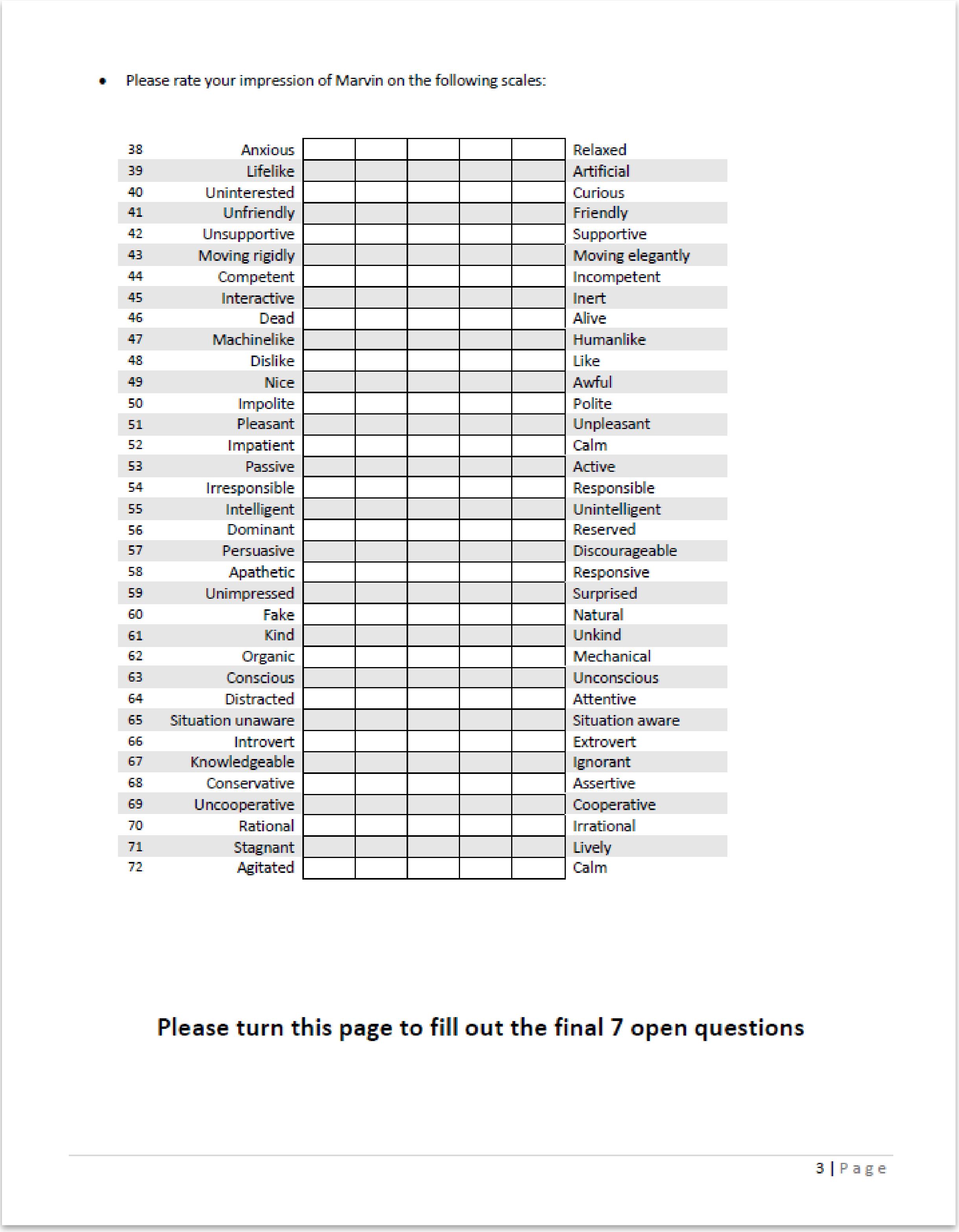
Post-Experiment Questions:

- The questions below refer to the robot by using its name ‘Marvin’.
- Mark every answer by crossing one of the check boxes.

Meaning of the check boxes:

Strongly disagree	disagree	somewhat disagree	neutral	somewhat agree	agree	strongly agree
- - -	- -	-	±	+	++	+++

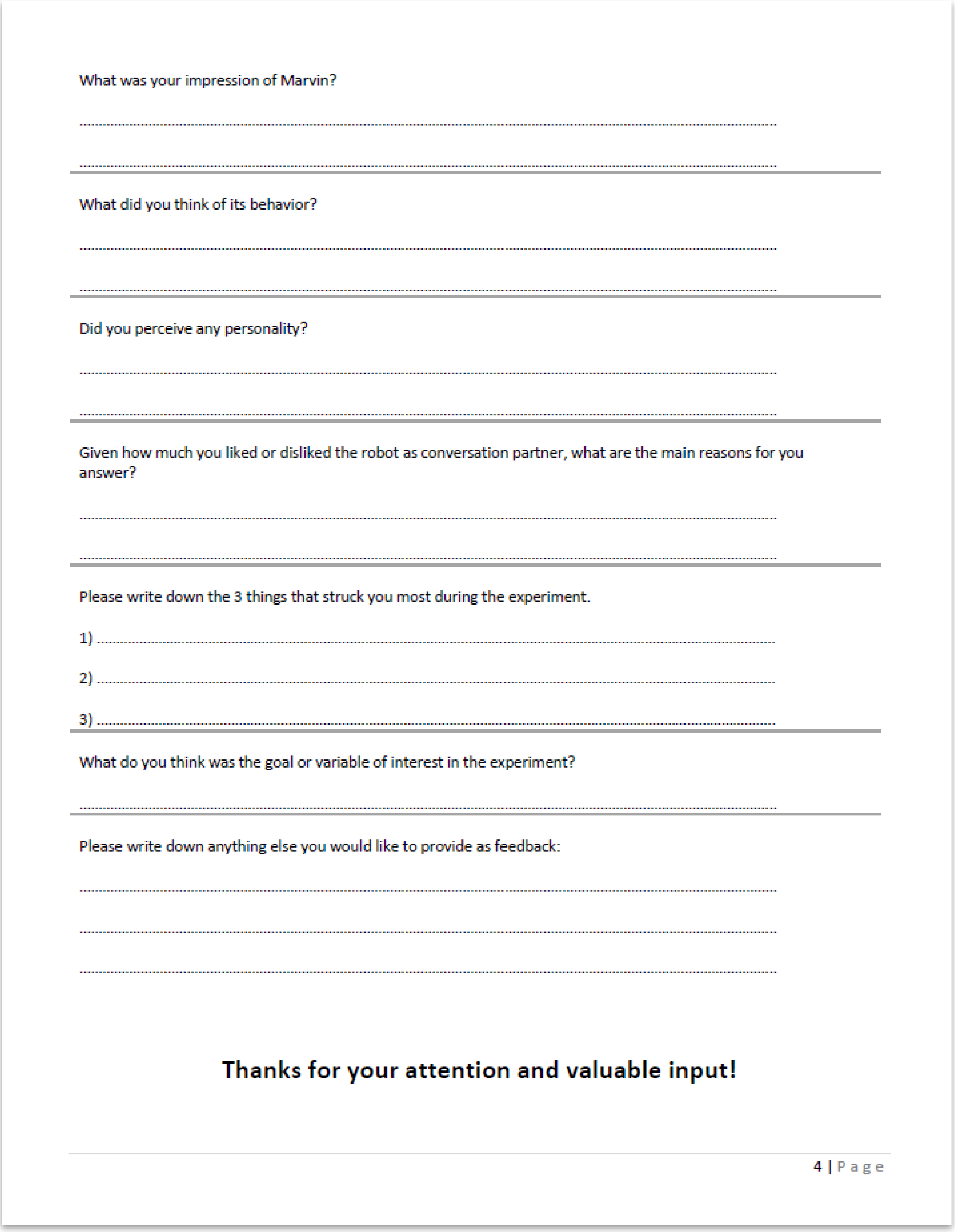
1	A high level of concentration is required when using Marvin	- - -	- -	-	±	+	++	+++
2	I felt tense using Marvin	- - -	- -	-	±	+	++	+++
3	Marvin didn't always do what I wanted	- - -	- -	-	±	+	++	+++
4	I would use this system	- - -	- -	-	±	+	++	+++
5	The interaction with Marvin is frustrating	- - -	- -	-	±	+	++	+++
6	It took Marvin too long to respond to my statements	- - -	- -	-	±	+	++	+++
7	The interaction with Marvin is repetitive	- - -	- -	-	±	+	++	+++
8	Marvin didn't always do what I expected	- - -	- -	-	±	+	++	+++
9	Marvin is accurate	- - -	- -	-	±	+	++	+++
10	I felt confident using Marvin	- - -	- -	-	±	+	++	+++
11	I knew what I could say or do at each point in the conversation with Marvin	- - -	- -	-	±	+	++	+++
12	I felt calm using Marvin	- - -	- -	-	±	+	++	+++
13	Marvin makes few errors	- - -	- -	-	±	+	++	+++
14	Marvin is friendly	- - -	- -	-	±	+	++	+++
15	Marvin is easy to use	- - -	- -	-	±	+	++	+++
16	I sometimes wondered if I was using the right word	- - -	- -	-	±	+	++	+++
17	I was not always sure what Marvin was doing	- - -	- -	-	±	+	++	+++
18	I felt in control of the interaction with Marvin	- - -	- -	-	±	+	++	+++
19	Marvin is unreliable	- - -	- -	-	±	+	++	+++
20	The interaction with Marvin is fast	- - -	- -	-	±	+	++	+++
21	I was able to recover easily from errors	- - -	- -	-	±	+	++	+++
22	Marvin is useful	- - -	- -	-	±	+	++	+++
23	Marvin responds too slowly	- - -	- -	-	±	+	++	+++
24	I enjoyed using Marvin	- - -	- -	-	±	+	++	+++
25	Marvin is pleasant	- - -	- -	-	±	+	++	+++
26	The interaction with Marvin is boring	- - -	- -	-	±	+	++	+++
27	I always knew what to say to Marvin	- - -	- -	-	±	+	++	+++
28	The interaction with Marvin is consistent	- - -	- -	-	±	+	++	+++
29	Marvin is stubborn	- - -	- -	-	±	+	++	+++
30	It is clear how to speak to Marvin	- - -	- -	-	±	+	++	+++
31	The interaction with Marvin is irritating	- - -	- -	-	±	+	++	+++
32	It is easy to lose track of where you are in an interaction with Marvin	- - -	- -	-	±	+	++	+++
33	The interaction with Marvin is efficient	- - -	- -	-	±	+	++	+++
34	The interaction with Marvin is unpredictable	- - -	- -	-	±	+	++	+++
35	It is easy to learn to use Marvin	- - -	- -	-	±	+	++	+++
36	Marvin is dependable	- - -	- -	-	±	+	++	+++
37	Marvin responded quickly	- - -	- -	-	±	+	++	+++



- Please rate your impression of Marvin on the following scales:

38	Anxious						Relaxed
39	Lifelike						Artificial
40	Uninterested						Curious
41	Unfriendly						Friendly
42	Unsupportive						Supportive
43	Moving rigidly						Moving elegantly
44	Competent						Incompetent
45	Interactive						Inert
46	Dead						Alive
47	Machinelike						Humanlike
48	Dislike						Like
49	Nice						Awful
50	Impolite						Polite
51	Pleasant						Unpleasant
52	Impatient						Calm
53	Passive						Active
54	Irresponsible						Responsible
55	Intelligent						Unintelligent
56	Dominant						Reserved
57	Persuasive						Discourageable
58	Apathetic						Responsive
59	Unimpressed						Surprised
60	Fake						Natural
61	Kind						Unkind
62	Organic						Mechanical
63	Conscious						Unconscious
64	Distracted						Attentive
65	Situation unaware						Situation aware
66	Introvert						Extrovert
67	Knowledgeable						Ignorant
68	Conservative						Assertive
69	Uncooperative						Cooperative
70	Rational						Irrational
71	Stagnant						Lively
72	Agitated						Calm

Please turn this page to fill out the final 7 open questions



What was your impression of Marvin?

....

....

What did you think of its behavior?

....

....

Did you perceive any personality?

....

....

Given how much you liked or disliked the robot as conversation partner, what are the main reasons for you answer?

....

....

Please write down the 3 things that struck you most during the experiment.

1) ....

2) ....

3) ....

What do you think was the goal or variable of interest in the experiment?

....

Please write down anything else you would like to provide as feedback:

....

....

....

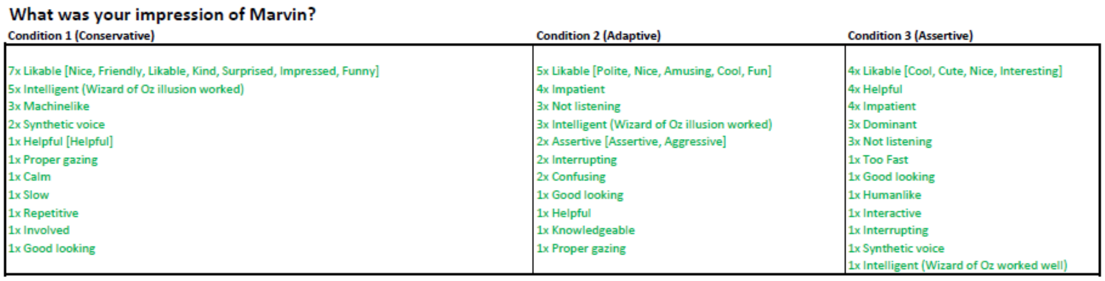
Thanks for your attention and valuable input!

APPENDIX B: Axial coding example and answer frequencies result

Participants' exact answers:	Nearest categorical equivalents:
- Sometimes the comments in between questions weren't entirely made at the right moment, but mostly it was pleasant	Awkward comment timings, pleasant
- Active	Active
- Fine, in the beginning you have to get used to it of course, but in the end it seems like a nice person	Fine, Have to get used to it, nice
- His head movement was a strong cue and felt natural, when speaking it really looked like it paid attention the head turning took some time, but he regularly looked at us	Natural head movements, pays attention Looks at you
- Seemed natural.	Natural
- He really looked at you when asking questions and after that he could also ask the other person.	Looked at you, Included both members
- Spontaneous, more than expected, and quite lively	Spontaneous, lively
- He still seems very awkward, but conversation goes smoothly, sometimes he responses quite well	Awkward, Smooth conversation
- His head movement was like humans in a conversation	Natural head movements
- He could have asked more open questions	Limited open questions
- Behaviour was close to that of a human, except that communication was almost one-directional	Humanlike, one directional communication
- He behaved quite professional and helpful. He managed to look at both of us and asked us both things	Helpful, Looked at you, Included both members
- He was nice, didn't interrupt me	Friendly, no interruptions
- Friendly and nice	Friendly

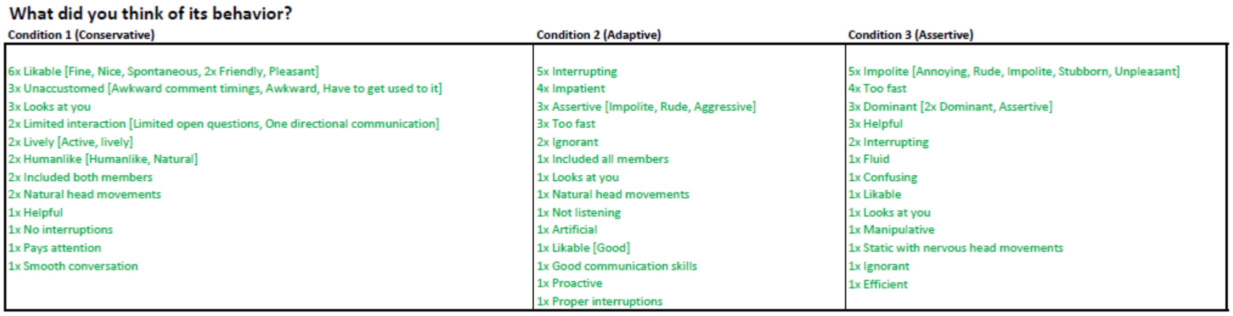
Sorted frequency list:

6x Likable [Fine, Nice, Spontaneous, 2x Friendly, Pleasant]
3x Unaccustomed [Awkward comment timings, Awkward, Have to get used to it]
3x Looks at you
2x Limited interaction [Limited open questions, One directional communication]
2x Lively [Active, lively]
2x Humanlike [Humanlike, Natural]
2x Included both members
2x Natural head movements
1x Helpful
1x No interruptions
1x Pays attention
1x Smooth conversation



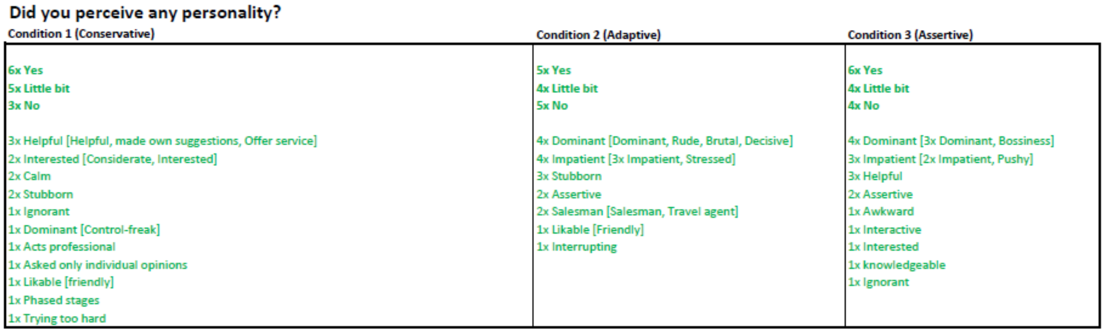
What was your impression of Marvin?

Condition 1 (Conservative)	Condition 2 (Adaptive)	Condition 3 (Assertive)
7x Likable [Nice, Friendly, Likable, Kind, Surprised, Impressed, Funny]	5x Likable [Polite, Nice, Amusing, Cool, Fun]	4x Likable [Cool, Cute, Nice, Interesting]
5x Intelligent (Wizard of Oz illusion worked)	4x Impatient	4x Helpful
3x Machinelike	3x Not listening	4x Impatient
2x Synthetic voice	3x Intelligent (Wizard of Oz illusion worked)	3x Dominant
1x Helpful [Helpful]	2x Assertive [Assertive, Aggressive]	3x Not listening
1x Proper gazing	2x Interrupting	1x Too Fast
1x Calm	2x Confusing	1x Good looking
1x Slow	1x Good looking	1x Humanlike
1x Repetitive	1x Helpful	1x Interactive
1x Involved	1x Knowledgeable	1x Interrupting
1x Good looking	1x Proper gazing	1x Synthetic voice
		1x Intelligent (Wizard of Oz worked well)



What did you think of its behavior?

Condition 1 (Conservative)	Condition 2 (Adaptive)	Condition 3 (Assertive)
6x Likable [Fine, Nice, Spontaneous, 2x Friendly, Pleasant]	5x Interrupting	5x Impolite [Annoying, Rude, Impolite, Stubborn, Unpleasant]
3x Unaccustomed [Awkward comment timings, Awkward, Have to get used to it]	4x Impatient	4x Too fast
3x Looks at you	3x Assertive [Impolite, Rude, Aggressive]	3x Dominant [2x Dominant, Assertive]
2x Limited interaction [Limited open questions, One directional communication]	3x Too fast	3x Helpful
2x Lively [Active, lively]	2x Ignorant	2x Interrupting
2x Humanlike [Humanlike, Natural]	1x Included all members	1x Fluid
2x Included both members	1x Looks at you	1x Confusing
2x Natural head movements	1x Natural head movements	1x Likable
1x Helpful	1x Not listening	1x Looks at you
1x No interruptions	1x Artificial	1x Manipulative
1x Pays attention	1x Likable [Good]	1x Static with nervous head movements
1x Smooth conversation	1x Good communication skills	1x Ignorant
	1x Proactive	1x Efficient
	1x Proper interruptions	



Did you perceive any personality?

Condition 1 (Conservative)	Condition 2 (Adaptive)	Condition 3 (Assertive)
6x Yes	**5x Yes**	**6x Yes**
5x Little bit	**4x Little bit**	**4x Little bit**
3x No	**5x No**	**4x No**
3x Helpful [Helpful, made own suggestions, Offer service]	4x Dominant [Dominant, Rude, Brutal, Decisive]	4x Dominant [3x Dominant, Bossiness]
2x Interested [Considerate, Interested]	4x Impatient [3x Impatient, Stressed]	3x Impatient [2x Impatient, Pushy]
2x Calm	3x Stubborn	3x Helpful
2x Stubborn	2x Assertive	2x Assertive
1x Ignorant	2x Salesman [Salesman, Travel agent]	1x Awkward
1x Dominant [Control-freak]	1x Likable [Friendly]	1x Interactive
1x Acts professional	1x Interrupting	1x Interested
1x Asked only individual opinions		1x knowledgeable
1x Likable [friendly]		1x Ignorant
1x Phased stages		
1x Trying too hard		

APPENDIX C: Dialog manager constructs

Construct name		Description
textline	=	Arbitrary utterance that is not part of the Dialog Manager (DM), used to verbalize sentences independently from the DM
introduction	=	Large text fragment containing a concise version of the experiment informed consent form
topicoptions	=	Utterances that introduce a new holiday topic (new stage) together with choices participants can choose from
alt1_topicoptions	=	When the number of possible choices are large, the set of options is split in half, and covered separately to avoid choice overload
holiday topic sequence *(11 dialog stages):*	=	This is the backbone of the dialog manager and represents the conversation possibilities during each experiment session. For the current study a holiday travel agency topic is chosen.
1. START **2. CONTINENT** **3. CITY** **4. PERIOD** **5. ACCOMODATION** **6. DURATION** **7. MOBILITY** **8. HOLIDAYTYPE** **9. THINGSTODO** **10. BUDGET** **11. END**	=	1. First stage utterance part of the dialog manager, confirming the experiment has officially started 2. Containing 6 continents to choose from 3. Containing 5 cities to choose from that are located in the earlier chosen continent 4. Preferred travel season (i.e. 4 choices) 5. Type of accommodation preferred to spend the nights, 6 choices available 6. Total holiday duration ranging from a quick visit to a longer stay, 3 choices available 7. The means of transportation preferred once arrived at the holiday destination, 5 choices given 8. A rough classification of the holiday, providing 3 types to choose from 9. Depending on the chosen holiday type, 2 activities are suggested to choose from 10. A rough estimation of the available budget is asked, 4 budget categories are given 11. A summary of all the choices made is given to inform the participants, including a word of gratitude and the notice that the experiment will continue to the next phase
longturnindicator	=	Containing a very short utterance meant to show that the robot is still paying attention and, independently from any given turn take condition, will not be completely silent during a participant's turn
breaksilence	=	When participants stay silent for a certain time after the Nao has introduced a new holiday topic with corresponding options, these type of utterance encourage the participants to speak up and effectively break the silence to start a conversation
TT_NoneFirstOPT	=	Utterances that asks the participant's opinion about one specific choice option besides from the first already proposed option.
direct_turntake	=	Asking the current speaker for any additional information regarding his/her opinion on the subject matter
switch_turntake	=	Asking the currently silent participant for any additional information regarding his/her opinion on the subject matter
info	=	Utterances that give very specific often factual information about some choice option (79 unique sentences defined)
Prefix *(3 prefix categories):*	=	Utterances that are cascaded to the beginning of an 'info' utterance to provide a fitting introduction, which are matched to the opinion of the participants regarding the option that is referred to in the ‘info’ utterance. Without prefix the 'info' utterances seem misplaced and awkward. A total of 3 prefix categories are used:
1. ProPrefix **2. ConPrefix** **3. OpenPrefix**	=	1. Used when participants are already optimistic about the current option 2. Used when participants are already pessimistic about the current option 3. Used when participants are indecisive about the current option
SpeedUP	=	Utterance to ask participants for confirmation regarding mutual agreement on a preferred option (if true the current dialog stage will be prematurely finished)
VerdictUtt *(4 verdict categories):*	=	Utterances that are used to wrap up the current dialog stage while securing a single option from the set provided. Depending on the course of the dialog however, the robot uses one of the following 4 utterance categories:
1. Multiple **2. Single** **3. Unknown** **4. None**	=	1. Several choices where preferred by the participants, therefor the DM will randomly choose one of the set of suitable options 2. One option was preferred, therefor the Nao will simply provide the participants with confirmation on their choice 3. The preferred option remains unknown (no agreement was reached), therefor the DM will randomly choose one option from the entire set 4. All options were rejected, therefor the DM will choose (from a set of non-mentioned options) a spare option to continue with.

APPENDIX D: Scripted robot utterances

```
textline = ("Welcome to the use lab and our little discussion group, I will explain more to you in a short moment, "+
            "but before we introduce ourselves I would like you to be absolutely quiet for 2 seconds.")
textline = "Hi there, my name is Marvin. Who are you, only look at me while briefly introducing yourself!"
textline = "For calibration reasons, please only look at me from a fixed position and try not to move your head"
textline = "Welcome "+NameRight+", nice to meet you."
textline = "You also welcome of course! Please introduce yourself, and remember only to look at me while talking!"
textline = "Nice to meet you too "+NameLeft+"!"

introduction = ["As described in the informed consent, the current experiment you are participating in will be "+
                "a team-building exercise. The main goal is to figure out what a holiday should be like if you "+
                "have to travel and spend the entire vacation together. It is all about collaboration and "+
                "communication to figure out each others preferences and find a satisfying arrangement. "+
                "Everybody is expected to be actively involved in the discussion to reach a unanimous decision "+
                "for a holiday planning. Since the two of you are going on a hypothetical holiday together, "+
                "ask for each others opinion. Make sure to give appropriate arguments, whether you "+
                "feel positive, negative or even unsure about some option. "+
                "I will guide you in the choices to make and sometimes provide extra information about options, "+
                "but I cannot answer any questions for you! It is of importance that you do not move around "+
                "or change from position in this room until the experiment is finished. "+
                "If you have any questions or requests, this is the time to ask."]

textline = "Everything clear and ready to start "+NameRight+"?"
textline = "Everything clear and ready for you also "+NameLeft+"?"

starting = ["I am also ready, so let us start to talk and discuss your holiday options together."]

# Dialog topic sequence:
#0 START
#1 CONTINENT
#2 CITY
#3 PERIOD
#4 ACCOMODATION
#5 DURATION
#6 MOBILITY
#7 HOLIDAYTYPE
#8 THINGSTODO
#9 BUDGET
#10 END

topicoptions = {
    0:starting,

    1:["What continent or worldpart would you prefer to visit, please discuss if it is going to be "]+
      [option[1]]+[", "]+[option[2]]+[", or "]+[option[3]]+["."],

    2:["What city in "]+[CONTINENT]+[" would you like to see, I have four options. "]+[option[1]]+[", "]+
      [option[2]]+[", "]+[option[3]]+[" and "]+[option[4]]+[". Let us first talk about "]+[option[1]]+[" and "]+
      [option[2]]+["?"],

    3:["During which time period of the year would you think is best to be in "]+[CITY]+[", in the "]+
      [option[1]]+[", "]+[option[2]]+[", "]+[option[3]]+[" or perhaps in the "]+[option[4]]+["?"],

    4:["Now we know the where, it is time to define the what. Let us start with the type of accommodation you would both prefer.
      "]+["Perhaps a budget solution like a "]+[option[1]]+[", "]+[option[2]]+[" or "]+[option[3]]+[" would be fine?"],

    5:["This is going great, only a few more steps are needed to complete your holiday definition. "]+
      ["Also important to determine is the total duration of your holiday stay. "]+
      ["Would you prefer to go "]+[option[1]]+[", " ]+[option[2]]+[", or rather "]+[option[3]]+["?"],

    6:["Taking the plane, you will fly to "]+[CITY]+[". Once you have arrived, you need to travel around and explore your
       surroundings. "]+["What transportation options are preferred, going by "]+[option[1]]+[" , "]+[option[2]]+
      [" , or rather by "]+[option[3]]+["?"],

    7:["While keeping in mind the travel period in "]+[PERIOD]+[", what type of vacation would you prefer? "]+
      ["An "]+[option[1]]+[", "]+[option[2]]+[" or "]+[option[3]]+[" vacation?"],

    8:["Given your choice to have an "]+[HOLIDAYTYPE]+[" vacation, what activity would definitely be on your too do list? "]+
      [option[1]]+[" or "]+[option[2]]+["?"],

    9:["And last but not least, it is good to get a mutual agreement on the total budget that has to be reserved. "]+
      ["Are you planning to spend "]+[option[1]]+[", or "]+[option[2]]+[", or willing to spend "]+[option[3]]+[", or even"]+
      [option[4]]+["?"],

    10:ending}

alt1_topicoptions = {2:["And regarding "]+[option[3]]+[" and "]+[option[4]]+["?"],
                     4:["Or rather go for a bit more comfort in a "]+[option[4]]+[", "]+[option[5]]+[" or "]+[option[6]]+["?"],
                     6:["And what about more healty alternatives such as going by "]+[option[4]]+[" or by "]+[option[5]]+["?"]}

continent = ["Asia",
             "Africa",
             "United States",
             "Central South America",
             "Europe",
             "Australia"]

city = {continent[0]: ["Singapore", "Hong Kong", "Bali", "Tokyo", "Maldives"],
        continent[1]: ["Cape Town", "Cairo", "Marrakech", "Tanzania", "Seychelles"],
        continent[2]: ["San Francisco", "New York", "Las Vegas", "New Orleans", "Miami"],
        continent[3]: ["Buenos Aires", "Rio de Janeiro", "Argentine", "Santiago", "Costa Rica"],
        continent[4]: ["Berlin", "Paris", "Barcelona", "Rome", "Copenhagen"],
        continent[5]: ["Darwin", "Brisbane", "Sydney", "Melbourne", "Perth"]}
```

```
travelperiod = ["Summer",
                "Autumn",
                "Winter",
                "Spring"]

accomodation = ["Hostel",
                "Camping",
                "Bed and Breakfast",
                "Hotel",
                "Holiday resort",
                "Local residence"]

tripduration = ["a few days",
                "a few weeks",
                "a few months"]

mobility = ["Public transfer",
            "Hitchhiking",
            "Car",
            "Foot",
            "Bike"]

holidaytype = ["Active",
               "Relaxing",
               "Partying"]

thingstodo = {holidaytype[0]: ["Visiting famous architectural buildings", "Seeking adventure and entertainment"],
              holidaytype[1]: ["Explore green areas of local nature", "Simply enjoy good sunny weather"],
              holidaytype[2]: ["Simply enjoy good sunny weather", "Visit popular clubs and enjoy nightlife"]}

budget = ["less than 1000 euros",
          "up to 1500 euros",
          "up to 2000 euros",
          "more than 2000 euros"]

ending = [["Well thats about it. With all the information combined, you have arranged yourselves a holiday for "]+[DURATION]+
          [" that will bring you to the continent of "]+[CONTINENT]+[" in the city "]+[CITY]+[" , during the "]+[PERIOD]+[". "]+
          ["Once arrived you will have a typical "]+[HOLIDAYTYPE]+[" vacation primarily going to "]+[THINGSTODO]+[" while "]+
          ["spending the nights in a comfortable "]+[ACCOMODATION]+[". To explore your surroundings you will mainly go by "]+
          [MOBILITY]+[ ", which also partially influences the estimated necessary total budget that will require you to spend "]+
          [BUDGET]+[" on this holiday. "]+
          ["Thanks for having participated in our dialog, the experiment will now continue to the next phase."]]

longturnindicator = ["Interesting",
                     "I see",
                     "Ok",
                     "Indeed",
                     "Go on"]

breaksilence = ["So who has any ideas?",
                "So what do you both think?",
                "Who of you can say something about it?",
                "Let us try to share some ideas.",
                "There must be some opinion about it."]

TT_NoneFirstOPT = [["How about "]+[OPT]+["?"],
                   ["What about "]+[OPT]+["?"],
                   ["What do you think of "]+[OPT]+["?"],
                   ["What is your opinion about "]+[OPT]+["?"],
                   ["What can you say about "]+[OPT]+["?"],
                   ["Any other judgement about "]+[OPT]+["?"],
                   ["What is your mind on "]+[OPT]+["?"]]

direct_turntake = ["Why do you think that?",
                   "Why do you see it that way?",
                   "Please elaborate a bit more.",
                   "Any more comments?",
                   "What would be another reason?",
                   "And what other reason would there be?",
                   "What could be a counterargument?",
                   "What would be an opposing statement?",
                   "What would your conversation partner most likely think?",
                   ["What would "]+[OTHERNAME]+["s opinion be?"],
                   ["What do you think will be the opinion of "]+[OTHERNAME]+["?"],
                   ["Why could "]+[OTHERNAME]+[" think you are "]+[RIGHTWRONG]+["?"],
                   ["Why could your discussion partner think you are "]+[RIGHTWRONG]+["?"]]

switch_turntake = [["To what degree would this also be your opinion?"],
                   ["To what extent do you agree with "]+[CURRENTNAME]+["?"],
                   ["That is interesting, what about you?"],
                   ["Ok fair enough, how about you?"],
                   ["And what is your view?"],
                   ["How do you feel about that?"],
                   ["Anything to add regarding "]+[OPT]+["?"],
                   ["Given what "]+[CURRENTNAME]+[" said, how would you comment on that?"],
                   ["Given what "]+[CURRENTNAME]+[" said, what would be your opinion?"]]
```

```
info = {
    continent[0]:"Asias culture is rich in every sense, from heritage, architecture, and way of life, to the intense spirituality
                  of the people.",
    continent[1]:"Travelling to Africa offers many vacation options. From a boat tour and safari in the jungle, to a relaxing spa
                  and viewing the wildlife.",
    continent[2]:"The USA is a versatile land, which is not surprising when nearly 4800 kilometres separate people on the west
                  coast from those on the east coast.",
    continent[3]:"America is a super colourful continent, meaning that the people, their clothes, the music, and just life itself
                  there is very diverse.",
    continent[4]:"With 27 countries located within the European Union alone, Europe offers a big cultural variety of travel
                  experiences.",
    continent[5]:"With its vast and varied landscapes, unique wildlife, and white sand beaches, Australia is one of the most
                  interesting continents around.",

    city[continent[0]][0]:"You can enjoy both urban and natural attractions in the mega metropolis Singapore.",
    city[continent[0]][1]:"It is said that Hong Kong will no doubt surprise you, and that there is an inspiring view of the
                           Symphony of the Stars lightshow from the promenade.",
    city[continent[0]][2]:"No matter which resort in Bali you would choose, it will most likely boast a beautiful beach, an
                           exotic spa, and an array full of dining options.",
    city[continent[0]][3]:"No trip to Tokyo would be complete without visiting some of the Buddhist and Shinto temples and
                           shrines.",
    city[continent[0]][4]:"Despite the numerous options for things to do in the Maladives, most visitors simply lounge on the
                           palatial resort island of their choice.",

    city[continent[1]][0]:"Your could start a day in Cape Town with a morning trip up the Table Mountain from where you will be
                           able to enjoy spectacular views of the city.",
    city[continent[1]][1]:"Many visitors of Cairo go for a tour to the Pyramids of Giza, and see more of its ancient Egyptian
                           ruins.",
    city[continent[1]][2]:"If you like history you can spend most of your time in or around the Medina, Marrakechs fortified old
                           city.",
    city[continent[1]][3]:"Tanzania is mainly known for Serengeti National Park, which houses a huge population of wildlife large
                           mammals.",
    city[continent[1]][4]:"Famous for its white idyllic beaches, even the most popular stretches of sand in Seychelles are never
                           crowded.",

    city[continent[2]][0]:"The Golden Gate Bridge is a must see in San Francisco, just like a visit to Alcatraz Island to tour
                           the infamous federal prison.",
    city[continent[2]][1]:"You will be surprised by New Yorks flourishing art, night life scenes, and the many huge skyscrapers
                           and monuments.",
    city[continent[2]][2]:"A visit to Las Vegas will most likely revolve around the Strip, this is the place where you will find
                           all the iconic neon lights and famous sights.",
    city[continent[2]][3]:"Night-life and rolling good times are the main attractions in New Orleans, plentiful live music clubs
                           of nearly every style.",
    city[continent[2]][4]:"Relaxing at the beach is truly the best free activity possible in Miami.",

    city[continent[3]][0]:"Buenos Aires has much to offer like boutique-shopping, opera-watching, and tango-dancing.",
    city[continent[3]][1]:"If it is your first trip to Rio, you will want to savour a chilled coconut as you survey Copacabana
                           beach.",
    city[continent[3]][2]:"Whale watching and horseback riding are for the adventurous traveller ways you can get acquainted with
                           Argentine.",
    city[continent[3]][3]:"Impressive skyscrapers, colonial architecture and spectacular peaks all jockey for your attention in
                           Santiago.",
    city[continent[3]][4]:"Costa Ricas strikingly diverse terrain of forests, wildlife reserves, and tropical beaches, offers
                           something for every traveller.",

    city[continent[4]][0]:"Berlins history of battling ideologies makes for some of the most fascinating sightseeing in Europe.",
    city[continent[4]][1]:"If it is your first time to Paris, you will probably want to spend some time at the Eiffel Tower.",
    city[continent[4]][2]:"You do not want to miss out on seeing Gaudis La Sagrada Familia in Barcelona.",
    city[continent[4]][3]:"A must-see in ancient Rome on many travellers agenda is the Trevi Fountain.",
    city[continent[4]][4]:"You should definitely visit the Tivoli gardens in Copenhagen located nearby the Central Train
                           Station.",

    city[continent[5]][0]:"Quite fascinating to see in Darwin are the big termite mounds in Litchfield natural park.",
    city[continent[5]][1]:"If you are not afraid to get wet feet, maybe rent a kayak to paddle across the twisty river of
                           Brisbane.",
    city[continent[5]][2]:"In Sydney you should make time for the beach, Bondi and Coogee beach are favourites.",
    city[continent[5]][3]:"If you are a sports fan, visiting the Cricket Ground in Melbourne is essential.",
    city[continent[5]][4]:"Rottnest Island in Perth is a protected Class A nature reserve, perhaps nice to enjoy a little
                           nature.",

    travelperiod[0]+continent[0]:"In the summer, Asia is for a large part pretty hot, muggy, and typhoon-prone.",
    travelperiod[1]+continent[0]:"It is a good period to enjoy daytime temps of around thirty degrees with below average room
                                  rates in autumn.",
    travelperiod[2]+continent[0]:"While cool temperatures during winter will discourage some travellers, maybe you will actually
                                  think it is ok.",
    travelperiod[3]+continent[0]:"If you wish to avoid both winters climate and summers humidity, spring is an exceptional time
                                  to visit.",

    travelperiod[0]+continent[1]:"Spending summertime in a desert climate is not really advised for travellers.",
    travelperiod[1]+continent[1]:"Late fall marks a sweet spot in the tourism calendar, the summer heat retreats and the crowds
                                  have yet to arrive.",
    travelperiod[2]+continent[1]:"Winter is prime tourist season in Africa, with visitors hoping to pair sightseeing with
                                  pleasant weather.",
    travelperiod[3]+continent[1]:"Springtime is a great time to visit Africa since the winter crowds are waning and the weather
                                  is gorgeous.",

    travelperiod[0]+continent[2]:"People from all over the country are drawn by the hope for nice weather and the promise of
                                  summertime activities in autumn.",
    travelperiod[1]+continent[2]:"Fall marks a sweet spot for North Americas tourism. Believe it or not, the weather is often
                                  warmer now than it is in the summer.",
    travelperiod[2]+continent[2]:"If you do not mind the chilly winds, you will find that winter is a great time to spend in the
                                  United States.",
    travelperiod[3]+continent[2]:"You can beat the tourist rush by visiting the USA in the spring, when the weather is mild and
                                  hotel prices have yet to rise.",
    travelperiod[0]+continent[3]:"South America winter season is great if you want to meet more locals that enjoy the moderate
                                  weather",
    travelperiod[1]+continent[3]:"South America spring is an ideal time for seeking sun and adventure.",
    travelperiod[2]+continent[3]:"Peak season is autumn in South America, hotel prices can be inflated during these months.",
    travelperiod[3]+continent[3]:"Crowds and hot summer weather dissipate in May, but still expect high humidity.",

    travelperiod[0]+continent[4]:"Be aware that summer forms the tourist season with high temperatures, high humidity and high
                                  prices for everything.",
    travelperiod[1]+continent[4]:"In autumn tourist season slows and hotel rates fall a little bit while still having comfortable
                                  temperatures.",
    travelperiod[2]+continent[4]:"You will find some great deals if you travel during the winter season, but it will be a little
                                  chilly.",
    travelperiod[3]+continent[4]:"Spring season is possibly the ideal time to travel in Europe due to low prices and pleasant
                                  temperatures.",

    travelperiod[0]+continent[5]:"Although wintertime in Australia, do not let that label fool you since the calendar is filled
                                  with mostly sunny days.",
    travelperiod[1]+continent[5]:"While autumn season here, the springtime in Australia is marked by warm days and breezy nights
                                  with an occasional serious rainfall.",
    travelperiod[2]+continent[5]:"Australias wet, humid summer season comes with temperatures reaching up to thirty degrees.",
    travelperiod[3]+continent[5]:"There is no need to pack anything more than a light jacket if you visit Australia during
                                  autumn.",

    accomodation[0]:"A pure functional and economic friendly choice would be an Hostel.",
    accomodation[1]:"A camping often offers multiple ways to stay like in a tent, caravan, or in a bungalow.",
    accomodation[2]:"Bed and Breakfast is a perfect choice if you want to save money and enjoy an already served breakfast.",
    accomodation[3]:"A hotel is a sound choice if you aim for a more catered holiday.",
    accomodation[4]:"If you like the safe environment with extra leisure options, a holiday resort is a good option.",
    accomodation[5]:"For a more personal and cultural experience, staying in the residence home of a local offers a unique
                     chance.",

    tripduration[0]:"Sometimes shorter vacations make a more memorable experience.",
    tripduration[1]:"Going for a few weeks will allow for more extensive sightseeing.",
    tripduration[2]:"Going away for a few months can really change your perspective on things.",

    mobility[0]:"Public transport is often very comfortable and save because of low accident probabilities compared to other
                 means of transportation.",
    mobility[1]:"Hitchhiking is mainly preferred as an economic way of travelling that also offers a unique way of getting to
                 know the locals.",
    mobility[2]:"Going by car enables fast transportation for you to travel whenever and wherever you want.",
    mobility[3]:"Walking maybe slow, but creates lots of opportunities to see or stumble upon things that you would have
                 otherwise missed.",
    mobility[4]:"Cycling is great and cheap way to explore a wide area surrounding your accommodation.",

    thingstodo[holidaytype[0]][0]:"The design of prominent buildings are often a reflection of the era and culture in which they
                                   were built.",
    thingstodo[holidaytype[0]][1]:"It may cost a little bit more, but it is worth visiting big tourist attractions and
                                   activities.",
    thingstodo[holidaytype[1]][0]:"Just being out in nature has a positive effect on many peoples body, mind and soul.",
    thingstodo[holidaytype[1]][1]:"With the intention to enjoy good weather, keep in mind of course that predictions do not
                                   always become true.",
    thingstodo[holidaytype[2]][1]:"Whether you are into serious clubbing or just a quick dance, OPT is a good night out for one
                                   and all."}

ProPrefix = ["It is also nice to know that ",
             "I can assure you also that ",
             "Also interesting for you to know is that ",
             "Maybe you also like to hear that ",
             "Perhaps also good to know is that ",
             "A good fact to also know is that "]

ConPrefix = ["It may still be valuable to know that ",
             "I would still like to say that ",
             "Did you nevertheless know that ",
             "Were you aware of the fact that ",
             "Perhaps something to still consider is that ",
             "Did you also took into account that "]

OpenPrefix = ["I can tell you that ",
              "Interesting to know is that ",
              "Did you know that ",
              "According to my information, ",
              "It is often said that ",
              "I like to share with you that "]

SpeedUP = ["You both agree to the same choice?",
           "So we have a mutual agreement?",
           "I guess the answer is clear then?",
           "It is clear then what option it should be?",
           "It seems the answer is obvious then?",
           "No need to discuss more options I guess?"]
```

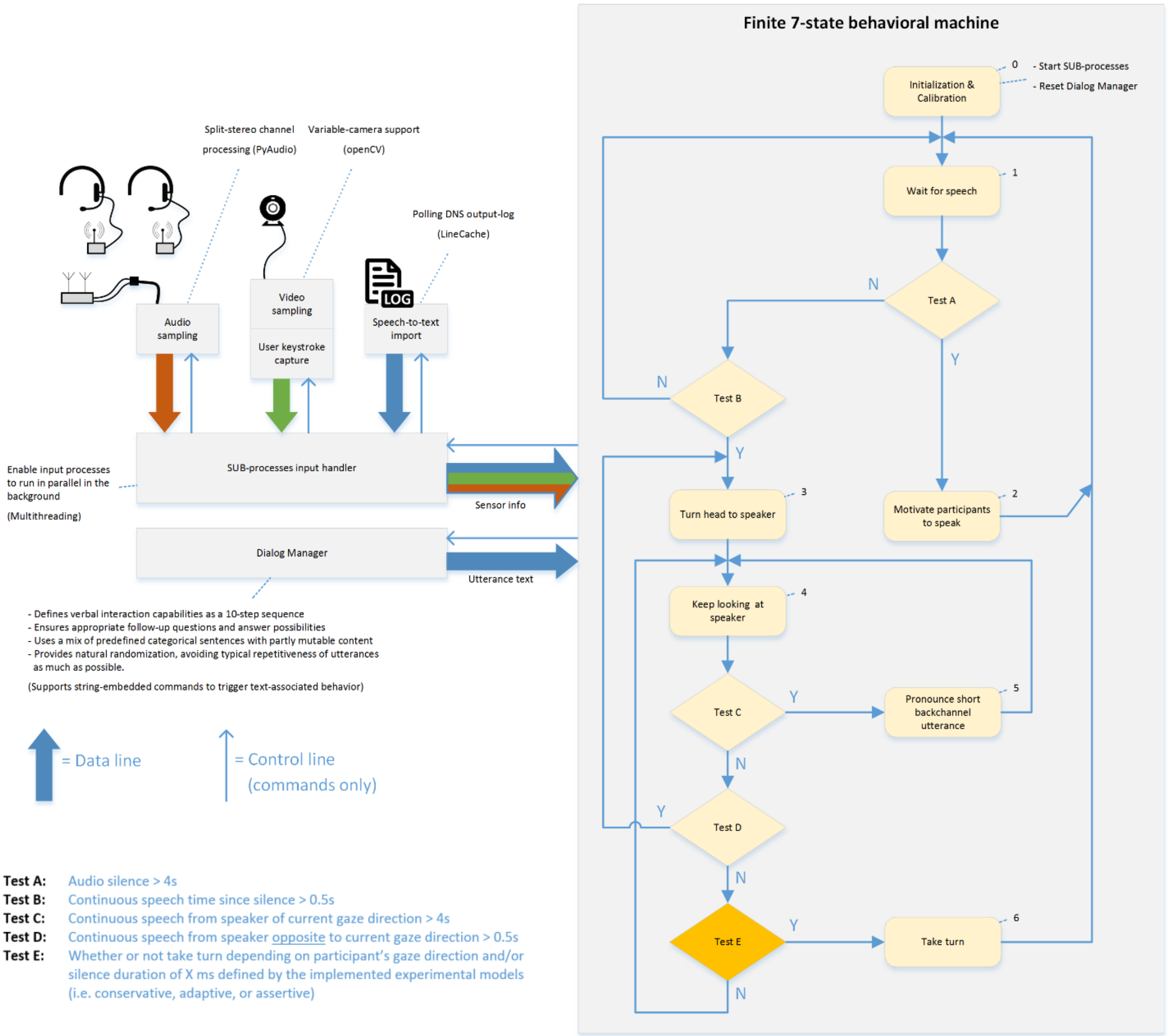
```
VerdictUtt = {
    "MULTIPLE": [ ["Well it seems that more than one option is possible, so for now let us go with "]+[OPT],
                ["You seem to are like minded, since there are multiple possibilities I will pick "]+[OPT]+[" for now "],
                ["Well with such positive judgements for multiple options, I will have to pick one which will be "]+[OPT]+['.'],
                ["I am glad you are both positive about several of my suggestions. To keep things simple, let us go with "]+
                [OPT]+['.'],
                ["Although it would not be impossible to combine several of the proposed options, for now we will stick with "]+
                [OPT]+['.']],

    "SINGLE": [ ["Clearly "]+[OPT]+[" seems to be the best choice."],
                ["I am glad that you found mutual agreement, "]+[OPT]+[" is selected for now."],
                ["Well this is an easy choice, I will note it is going to be "]+[OPT]+['.'],
                ["Ok, let me select option "]+[OPT]+[' for you'],
                ["It is clear to me that  "]+[OPT]+[' is preferred']],

    "UNKNOWN":[ ["Ok, not all arguments were clear to me, so I will choose for the time being "]+[OPT]+['.'],
                ["I am not completely sure what option would suit you best, let us just say for simplicity it will be "]+
                [OPT]+['.'],
                ["Unfortunately for me it was not one hundred percent clear what your preference is, so I will take a guess for
                "]+[OPT]+['.'],
                ["Maybe I was not listening close enough, but since no clear preference was given for one of the option, I will
                just go with "]+[OPT]+['.'],
                ["Well this time it looks like I am forced to make the decision for you. It will therefore be  "]+[OPT]+['.']],

    "NONE":   [ ["It is unfortunate that none of the options discussed so far seem to fit. "+
                 "Therefore I will choose an fall back option that we have maybe not mentioned earlier, which is "]+[OPT]+['.'],
                ["What a same that no options seems to suit your needs. "+
                 "Let me therefore choose an fall back option that we may not have mentioned earlier, namely "]+[OPT]+['.'],
                ["Well it seems that you are pretty picky and none of the suggestions was adequate enough. "+
                 "As fall back option, I will select "]+[OPT]+['.'],
                ["Did none of the options matched your thoughts? Well, then I shall for the time being just choose as fall back
                "]+[OPT]+['.'],
                ["Ok, although unanimously decided, rejecting all possibilities leaves me no choice other than to select
                "]+[OPT]+[" as fall back."]]}
```

APPENDIX E: System schematic and descriptions


Finite 7-state behavioral machine
Initialization & Calibration
0
- Start SUB-processes
- Reset Dialog Manager
Wait for speech
1
Test A
N
Y
Test B
N
Y
Turn head to speaker
3
Motivate participants to speak
2
Keep looking at speaker
4
Test C
Y
Pronounce short backchannel utterance
5
N
Test D
Y
N
Test E
Y
Take turn
6
N
Split-stereo channel processing (PyAudio)
Variable-camera support (openCV)
Polling DNS output-log (LineCache)
LOG
Audio sampling
Video sampling
User keystroke capture
Speech-to-text import
SUB-processes input handler
Sensor info
Enable input processes to run in parallel in the background
(Multithreading)
Dialog Manager
Utterance text
- Defines verbal interaction capabilities as a 10-step sequence
- Ensures appropriate follow-up questions and answer possibilities
- Uses a mix of predefined categorical sentences with partly mutable content
- Provides natural randomization, avoiding typical repetitiveness of utterances as much as possible.
(Supports string-embedded commands to trigger text-associated behavior)
= Data line
= Control line (commands only)

Test A: Audio silence > 4s
Test B: Continuous speech time since silence > 0.5s
Test C: Continuous speech from speaker of current gaze direction > 4s
Test D: Continuous speech from speaker opposite to current gaze direction > 0.5s
Test E: Whether or not take turn depending on participant's gaze direction and/or silence duration of X ms defined by the implemented experimental models (i.e. conservative, adaptive, or assertive)

Sensor info output (➡)

Variable name	*Description*	*Data type*
Iteration	Number of loops passed since main script has started	[integer]
Mic	Reflects stereo audio-channel activity with 4 unique symbols O, L, R, B for respectively silence, left, right, both	[char]
Facedetected	Boolean that tells if a person's face is recognized	[true/false]
Facecenter	Center coordinates (X,Y) a detected face	[float ; float]
Facesize	Width in pixels of a detected face	[integer]
Faceradians	Centre coordinates (X,Y) of a detected face in approx. radians	[float ; float]
Headpose	Head pose estimation (Yaw, Pitch) of a detected face	[float ; float]
DialogStats	Dialog statistics providing participant specific as well as overall data regarding the frequencies and lengths of turns, silences, turn-intervals, and turn-overlaps	[custom class]
KeyPressChar	Returns ASCII character which corresponds to keystroke by user	[char]
STTwords	Provides a list of strings representing the real-time interpretation of spoken utterances using the output-log of stand-alone voice recognition software	[list of strings]

APPENDIX F: System presets and control options

Audio script presets

Variable name	*Description*	*Default setting*
DEVICE_INDEX	Number to select audio recording device	Default = 0 (using MS Windows OS Auto Select)
SAMPLERATE	Audio channel(s) sampling rate	Default = 44100 (Hz)
INPUT_BLOCK_TIME	Audio sample length to be processed	Default = 0.02 (seconds)
CHANNELS	Number of channels to be sampled, mono or stereo	Default = “stereo” (string)
CreateAudioFile	Store audio recording as .wav file	Default = 0 (not enabled)
voice_threshold	RMS threshold for sound (speech) detection	Default = 0.01 (higher No. decreases sensitivity)
Intervalbuffer	Interval sensitivity, to ignore very short inter-speech silences	Default = 7 (floor((0.05/INPUT_BLOCK_TIME)*3))
Peakbuffer	Peak sensitivity, to ignore very short spikes of high RMS values	Default = 2 (floor((0.05/INPUT_BLOCK_TIME)*1))
WindowSizeSec=1	Moving average window size defined in seconds that is used to calculate dialog statistics	Default = 1 (s)
resetData	Used to reset the dialog statistics process	Default = 0 (no reset is performed)

Video script presets

Variable name	*Description*	*Default setting*
facedetectTYPE	Specifies type of face detection algorithm to be used: 1. Uses function defined in modified nao.py script 2. Uses robot build-in face detector from Aldebaran 3. Face detection using OpenCV Haar Cascades	Default = 1
faceYawPitchON	Audio channel(s) sampling rate	Default = 1 (head pose estimation is enabled)
Nao_IP	Audio sample length to be processed	Default = "192.168.0.114" (Nao robot IP-address)
indexEXTERNALcam	Number of channels to be sampled, mono or stereo	Default = 0 (using MS Windows OS Auto Select)
CreateVideoFile	Store video recording as .avi file	Default = 0 (not enabled) (Not supported for Nao robot build-in camera)
useNaocam	Specifies if the Nao robot build-in camera is used	Default = 1 (Nao camera is used)
dimension	Specifies camera capture resolution: kQQVGA (160x120) k4VGA (1280x720) kQVGA (320x240) k960p (1280x960) kVGA (640x480)	Default = kVGA (640x480)
maxFPS	Maximum allowed frames to capture per second	Default = 30 (often not even feasible)

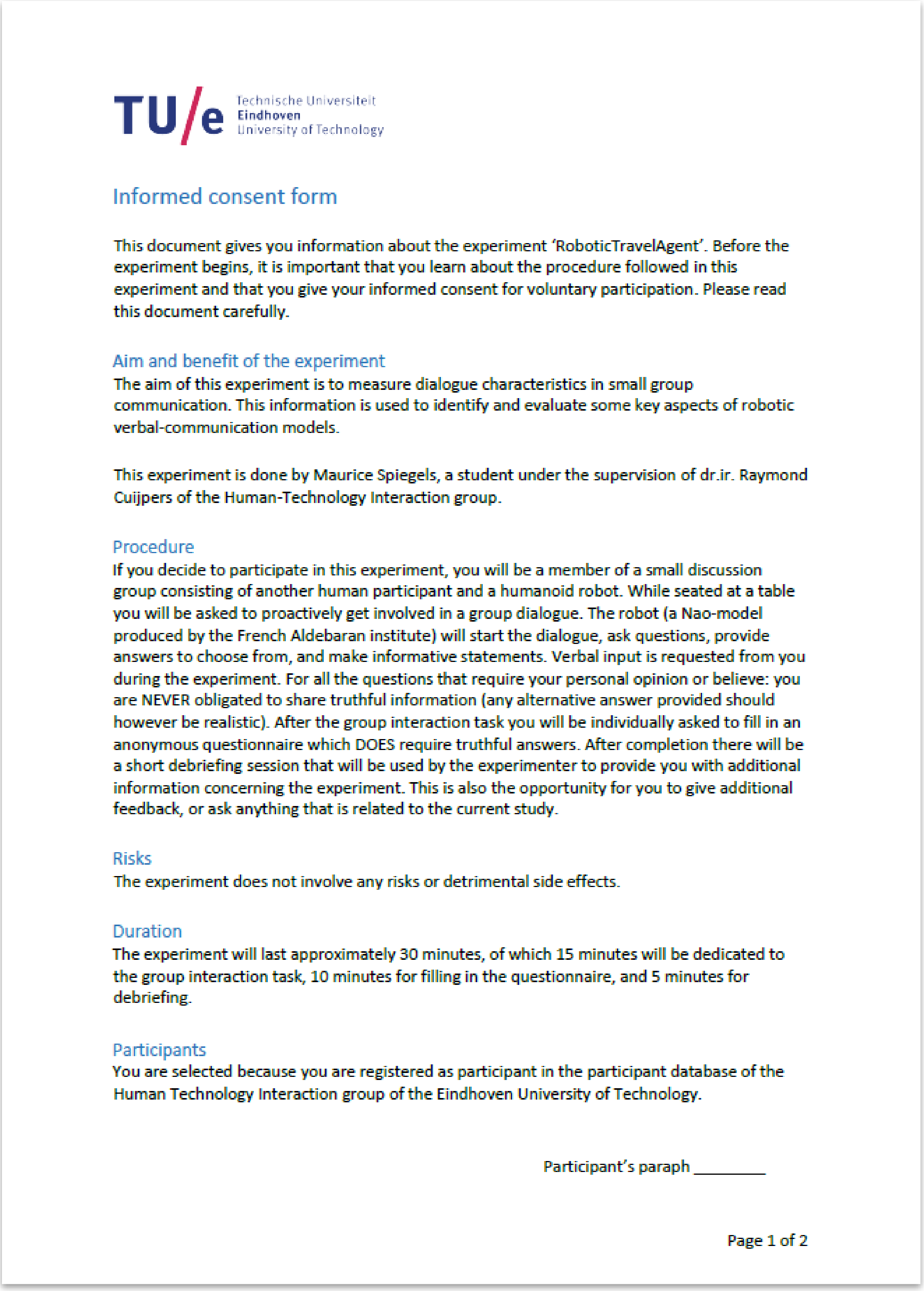
Speech-to-text script preset

Variable name	*Description*	*Default setting*
PollingRate	Inactivity time between successive checks for log file changes	Default = 0.01 (s)

Runtime experimenter manual control

Controls	*Function*
Keystroke P	pause behavior state machine (complete system freeze)
Keystroke ↑	adjust pitch + 0.2 (degrees)
Keystroke ↓	adjust pitch - 0.2 (degrees)
Keystroke →	adjust yaw + 0.2 (degrees)
Keystroke ←	adjust yaw - 0.2 (degrees)
Numpad [1-9]	select dialog answer option (provided by the dialog manager)
Numpad +	set selected option as positive (preferable by participants)
Numpad -	set selected option as negative (rejected by participants)
Numpad *	activate dialog speed-up to skip the remaining number of predefined turn takes for the current dialog stage
Numpad /	cancel (if any) dialog speed-up

APPENDIX G: Informed Consent



Technische Universiteit
Eindhoven
University of Technology

Informed consent form

This document gives you information about the experiment 'RoboticTravelAgent'. Before the experiment begins, it is important that you learn about the procedure followed in this experiment and that you give your informed consent for voluntary participation. Please read this document carefully.

Aim and benefit of the experiment

The aim of this experiment is to measure dialogue characteristics in small group communication. This information is used to identify and evaluate some key aspects of robotic verbal-communication models.

This experiment is done by Maurice Spiegels, a student under the supervision of dr.ir. Raymond Cuijpers of the Human-Technology Interaction group.

Procedure

If you decide to participate in this experiment, you will be a member of a small discussion group consisting of another human participant and a humanoid robot. While seated at a table you will be asked to proactively get involved in a group dialogue. The robot (a Nao-model produced by the French Aldebaran institute) will start the dialogue, ask questions, provide answers to choose from, and make informative statements. Verbal input is requested from you during the experiment. For all the questions that require your personal opinion or believe: you are NEVER obligated to share truthful information (any alternative answer provided should however be realistic). After the group interaction task you will be individually asked to fill in an anonymous questionnaire which DOES require truthful answers. After completion there will be a short debriefing session that will be used by the experimenter to provide you with additional information concerning the experiment. This is also the opportunity for you to give additional feedback, or ask anything that is related to the current study.

Risks

The experiment does not involve any risks or detrimental side effects.

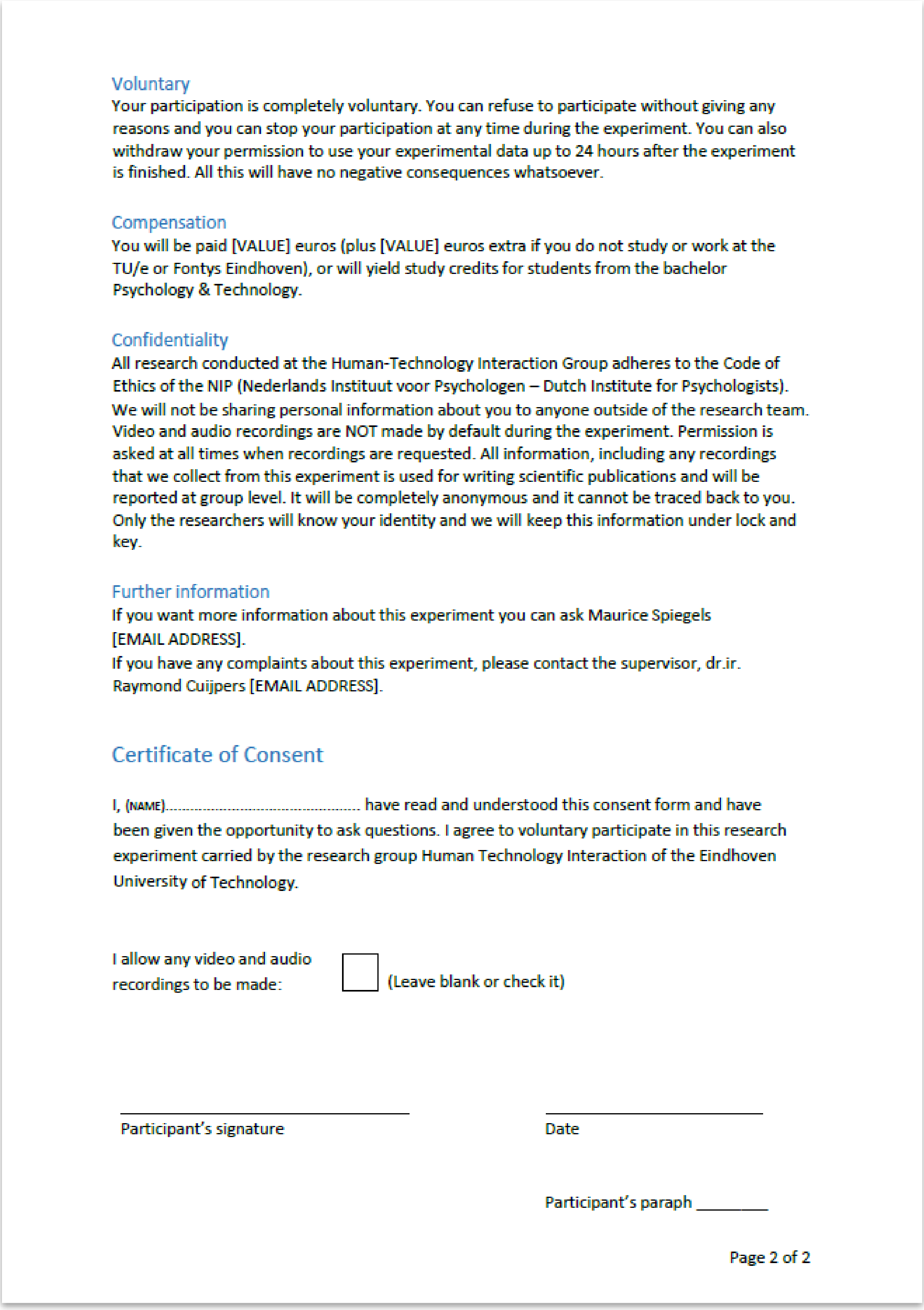
Duration

The experiment will last approximately 30 minutes, of which 15 minutes will be dedicated to the group interaction task, 10 minutes for filling in the questionnaire, and 5 minutes for debriefing.

Participants

You are selected because you are registered as participant in the participant database of the Human Technology Interaction group of the Eindhoven University of Technology.

Participant's paraph _________



Voluntary

Your participation is completely voluntary. You can refuse to participate without giving any reasons and you can stop your participation at any time during the experiment. You can also withdraw your permission to use your experimental data up to 24 hours after the experiment is finished. All this will have no negative consequences whatsoever.

Compensation

You will be paid [VALUE] euros (plus [VALUE] euros extra if you do not study or work at the TU/e or Fontys Eindhoven), or will yield study credits for students from the bachelor Psychology & Technology.

Confidentiality

All research conducted at the Human-Technology Interaction Group adheres to the Code of Ethics of the NIP (Nederlands Instituut voor Psychologen – Dutch Institute for Psychologists). We will not be sharing personal information about you to anyone outside of the research team. Video and audio recordings are NOT made by default during the experiment. Permission is asked at all times when recordings are requested. All information, including any recordings that we collect from this experiment is used for writing scientific publications and will be reported at group level. It will be completely anonymous and it cannot be traced back to you. Only the researchers will know your identity and we will keep this information under lock and key.

Further information

If you want more information about this experiment you can ask Maurice Spiegels [EMAIL ADDRESS].
If you have any complaints about this experiment, please contact the supervisor, dr.ir. Raymond Cuijpers [EMAIL ADDRESS].

Certificate of Consent

I, (NAME)... have read and understood this consent form and have been given the opportunity to ask questions. I agree to voluntary participate in this research experiment carried by the research group Human Technology Interaction of the Eindhoven University of Technology.

I allow any video and audio recordings to be made: □ (Leave blank or check it)

____
Participant's signature

____
Date

Participant's paraph ________

Page 2 of 2

APPENDIX H: Participant database and flyer invitation

Dear [NAME],

Already dreaming of a nice vacation to temporarily escape from everyday stress? Then this is a fun experience you should not miss! You are hereby invited to participate in the experiment 'RoboticTravelAgent'.

In this experiment you will be asked to join a small group conversation about your perfect holiday together with another participant and our robot. You should be able to speak and understand English. Afterwards you are required to fill in a single questionnaire. The experiment will last about 30 minutes.

The reimbursement is € [VALUE] with an additional € [VALUE] for participants from outside the TU/e or Fontys Eindhoven. Psychology & Technology students can receive course credits instead.

The experiment will take place in the IPO building at the TU/e campus. You will be expected at the laboratories hallway. Follow the signs "HTI laboratories" (entering the IPO building, immediately turn right, cross the hall, up a short stair, turn right again, and after 10 meters the hallway is on the left side). You will be picked up at the yellow seats in waiting-area a.

Since this experiment requires the presence of two participants simultaneously, please REGISTER FOR A TIMESLOT THAT ALREADY HAS A ONE RESERVATION, if available. If this is not convenient just subscribe for any other date and time. For the same reason it is very important to show up on time, and, if delayed, notify us.

You can register online using the following link:
[LINK]

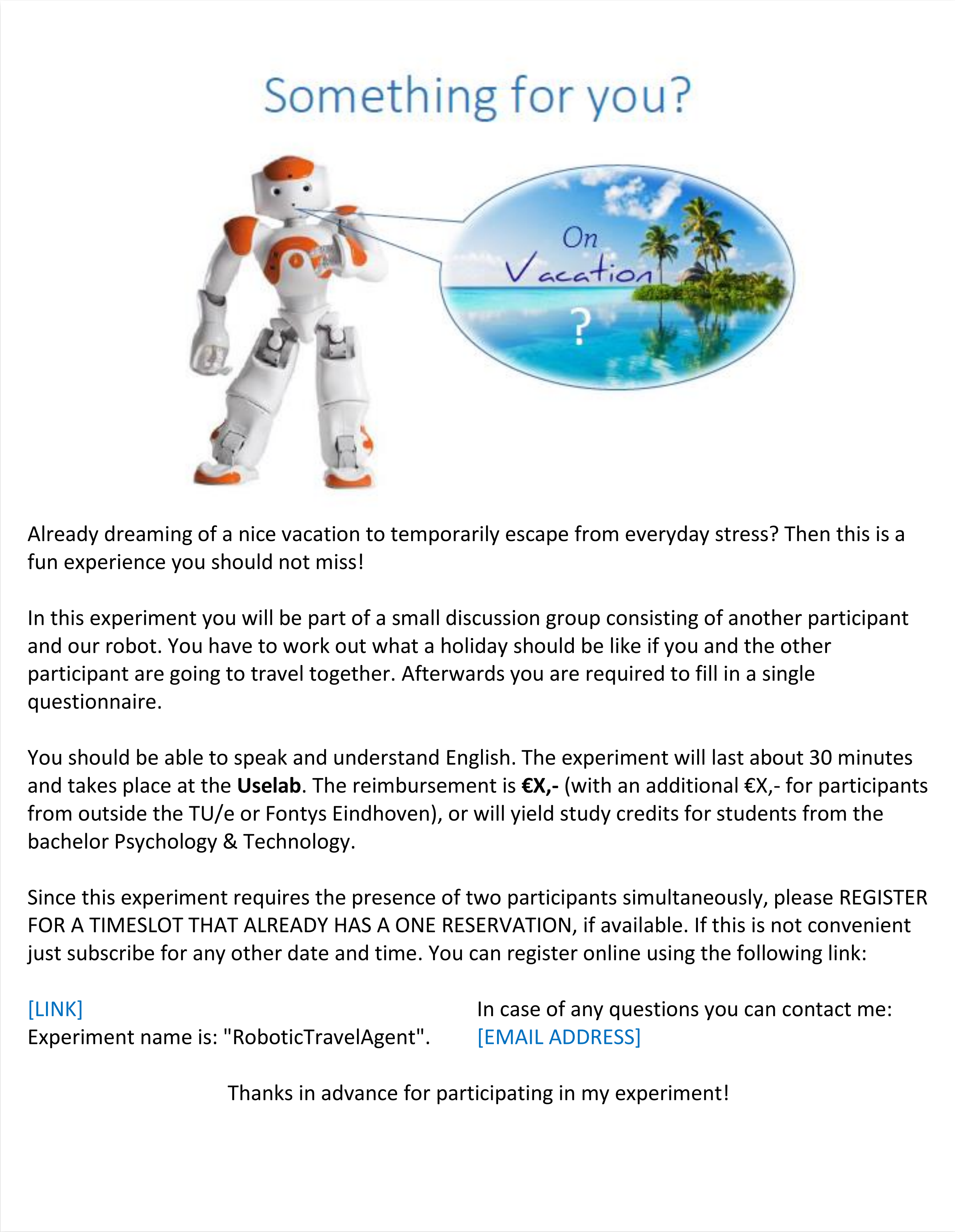
The name of the experiment is: "RoboticTravelAgent". Log in using your own e-mail address. You will get a reservation confirmation via email.

In case of any questions you can contact me:
[EMAIL ADDRESS], [MOBILE NUMBER]

Thanks in advance for your time and interest by participating in my experiment!

Kind regards,

[EXPERIMENTER NAME]


On
Vacation
?

Something for you?

Already dreaming of a nice vacation to temporarily escape from everyday stress? Then this is a fun experience you should not miss!

In this experiment you will be part of a small discussion group consisting of another participant and our robot. You have to work out what a holiday should be like if you and the other participant are going to travel together. Afterwards you are required to fill in a single questionnaire.

You should be able to speak and understand English. The experiment will last about 30 minutes and takes place at the **Uselab**. The reimbursement is **€X,-** (with an additional €X,- for participants from outside the TU/e or Fontys Eindhoven), or will yield study credits for students from the bachelor Psychology & Technology.

Since this experiment requires the presence of two participants simultaneously, please REGISTER FOR A TIMESLOT THAT ALREADY HAS A ONE RESERVATION, if available. If this is not convenient just subscribe for any other date and time. You can register online using the following link:

[LINK]
Experiment name is: "RoboticTravelAgent".

In case of any questions you can contact me:
[EMAIL ADDRESS]

Thanks in advance for participating in my experiment!