Tutoring Robots

Samer Al Moubayed, Jonas Beskow, Bajibabu Bollepalli, Ahmed Hussen-Abdelaziz, Martin Johansson, Maria Koutsombogera, José David Lopes, Jekaterina Novikova, Catharine Oertel, Gabriel Skantze, et al.

Tutoring Robots:

Multiparty Multimodal Social Dialogue with an Embodied Tutor

Samer Al Moubayed[1], Jonas Beskow[1], Bajibabu Bollepalli[1], Ahmed Hussen-Abdelaziz[5], Martin Johansson[1], Maria Koutsombogera[2], José David Lopes[3], Jekaterina Novikova[4], Catharine Oertel[1], Gabriel Skantze[1], Kalin Stefanov[1], and Gül Varol[6]

[1] KTH Speech, Music and Hearing, Sweden
[2] Institute for Language and Speech Processing- "Athena" R.C., Greece
[3] Spoken Language Systems Laboratory, INESC ID Lisboa, Portugal
[4] Department of Computer Science, University of Bath, UK
[5] Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany
[6] Department of Computer Engineering, Boğaziçi University, Turkey

sameram@kth.se, beskow@kth.se, bajibabu@kth.se, ahmed.hussenabdelaziz@rub.de, vhmj@kth.se, mkouts@ilsp.athena-innovation.gr, zedavid@l2f.inesc-id.pt, j.novikova@bath.ac.uk, catha@kth.se, skantze@kth.se, kalins@kth.se, gul.varol@boun.edu.tr

Abstract. This project explores a novel experimental setup towards building spoken, multi-modally rich, and human-like multiparty tutoring agent. A setup is developed and a corpus is collected that targets the development of a dialogue system platform to explore verbal and nonverbal tutoring strategies in multiparty spoken interactions with embodied agents. The dialogue task is centered on two participants involved in a dialogue aiming to solve a card-ordering game. With the participants sits a tutor that helps the participants perform the task and organizes and balances their interaction. Different multimodal signals captured and auto-synchronized by different audio-visual capture technologies were coupled with manual annotations to build a situated model of the interaction based on the participants personalities, their temporally-changing state of attention, their conversational engagement and verbal dominance, and the way these are correlated with the verbal and visual feedback, turn-management, and conversation regulatory actions generated by the tutor. At the end of this chapter we discuss the potential areas of research and developments this work opens and some of the challenges that lie in the road ahead.

Keywords: Multiparty, Multimodal, Turn-taking, Tutor, Conversational Dominance, Non-verbal Signals, Visual Attention, Spoken Dialogue, Embodied Agent, Social Robot.

adfa, p. 1, 2013.

1 Introduction

Today, advanced, reliable and real-time capture devices and modeling techniques are maturing and becoming significantly more accessible to researchers. Along with that, new findings in human-human conversations shed more light on the importance of modeling all the available verbal and non-verbal actions in conversations (in addition to the stream of words) and on how these are required in order to build more human-like dialogue systems that can be used by avatars and robots to exhibit natural behaviors (e.g. [1,2]). With these developments, research has been moving towards analyzing multiparty, multimodal conversations with the aim of understanding and modeling the structure and strategies with which interlocutors regulate the interaction, and keep their conversations rich, fluent, and successful.

Building socially aware and affective spoken dialogue systems has the potential of not only providing a hands-free interface for information input and output, but perhaps even more importantly, in many applications, the ability of using speech to provide a human-like interface that can understand and communicate all the subtle non-verbal signals that accompany the stream of sounds and provide significant information about the state of the user and the interpretation of the users verbal actions. These signals become even more central in scenarios where affective and social skills are essential for the success of the interaction (such as learning, collaborative task solving, games, and commerce [3-5]). Although the challenges and potentials of such social and affective technology are far from explored and understood, thanks to the recent availability and robustness of capture devices (e.g. microphone arrays, depth sensors), modeling techniques (e.g. speech recognizers, face tracking, dialogue modeling), and flexible and human like synthesis devices (e.g. avatars and humanoid robots), several recent projects are targeting the potential of different applications and high-end effects of modeling social and affective spoken interactions (e.g. Collaborative task solving in [6]; Education in [7]; Child therapy in [8]).

One major obstacle in the face of exploring the effects of spoken social and affective behavior of artificial embodied entities lies in the multidisciplinary nature of these setups and in the limitations of the different technologies that they involve. For example, while these applications aim at stimulating natural, fluent and spontaneous spoken behavior from the users, yet Automatic Speech Recognition systems (ASRs) still suffer a very limited power in handling such spoken conversational utterances, acoustically and grammatically. Another important challenge is how to keep these setups noninvasive, without hindering the fluency and spontaneity of the interaction (avoiding the use of cables, headsets, gaze trackers that dictate very little movement space in order for them to robustly function, etc.).

In this project, we target the development of a relatively natural, spoken, spatially and socially aware embodied talking head paying special attention to the aforementioned criteria.

The experimental design in this project is targeted towards multiparty collaborative task-solving, a research application that we expect to be central to the use of these technologies in the future. Such an application area is also rich with non-verbal and conversational variables that go beyond the meaning of words the users are using, but

extends to measuring other variables that play an important role in the interaction strategies and regulatory actions the agent should take into account, such as attention and conversational dominance.

2 Overview: The Moon-Survival Multiparty Tutor

Our work attempts to address interactional skills required by an embodied dialogue system to control the interaction flow as well as to boost and balance the engagement of the participants in the task they are involved in, while at the same time mitigating dominant behavior and encouraging less involved interlocutors to equally participate in the interaction. The task and the setup chosen in this work are considered as first steps towards understanding the behavior of a conversational tutor in multiparty task solving setups, as an example of a setup that can be used for applications in group-collaboration and negotiations, an activity that is highly dependent on the affective, and social behavior of the interlocutors [3]. Another main criterion that is taken into account when developing this setup is the ability to move directly from the models learnt from the annotations and analysis of the corpus, into an implementation of multiparty multimodal dialogue system, using the robot head Furhat [9], and the newly developed IrisTK dialogue platform [10] both developed and utilized in multimodal multiparty embodied spoken dialogue systems.

The interaction setup in this work consisted of two users and one tutor sitting around a round table. The two users' task is to discuss and negotiate the importance of certain objects and arrive to a decision on ordering them in terms of priority. The task was based on a shortened version of a "*NASA Exercise: Survival on the Moon*". During this exercise participants have to imagine that they are members of a space crew originally scheduled to rendezvous with a mother ship on the lit surface of the moon. However, due to mechanical difficulties, their ship was forced to land at a spot some 200 miles from the rendezvous point. During reentry and landing, much of the equipment aboard was damaged and, since survival depends on reaching the mother ship, the most critical items available must be chosen for the 200-mile trip. Two participants were presented with six cards with the pictures of six items left intact and undamaged after landing, as shown in Figure 2.

The tutor's task was to present the game, control its flow, and guarantee a high and balanced level of involvement between the two users and a collaborative decision process regarding the importance of the cards.

The remaining of the book chapter describes the project design and implementation that was done during the eNTERFACE2013 Workshop, that utilizes the Moon Survival Setup, as well as the requirements and development of the different technologies required to build a completely autonomous embodied dialogue system that could play the role of the tutor in similar setups. We firstly present a corpus collection study of a human-human setup, the design decision taken to reflect some of the affective and higher level conversational features that are present in collaborative task-solving, and outline some preliminary analysis of the data. After that, we describe a dialogue system setup where the human-tutor is replaced with the Furhat embodied talking head.

We also discuss the limitations of the work done, and possibilities for future research in this young and challenging area.

3 Experimental Setup

- *Physical setup:* Consists of a tutor and two participants, sitting at a round table, shaping an equilateral triangle.
- *Visual tracking:* Both subjects are tracked using two Kinect[1] sensors. The sensors are intended to capture the head-pose, facial expressions, and skeletal movement of both interlocutors. The Kinect sensors are placed at about 1.5 meter distance and outside the space of the interaction to limit the interference of the sensor on the interaction.
- *Auditory tracking:* Instead of using head-held close-range microphones (commonly used in dialogue recordings to limit the influence of overlapping speech); in this recording we took advantage of the Microcone™[2] multichannel microphone array - by Dev-Audio. Microcone™ consists of 6 channel microphone (over 360 degrees) that provides high quality far-field speech input, along with activation values for the different microphones, allowing for the detection of multiple speakers, and hence overlaps and speakers locations. Microcone™ was designed for automatic annotation of roundtable meetings and it provides a measure of the microphone activity at 20fps. This means that the device is able to infer the speakers location (even in cases of overlap), and would provide raw audio signal for all six microphones with a reliable beam-forming and noise suppression. The choice of a table-top microphone array over a headset is made for two reasons: if people eventually address a dialogue system using a headset, the speech might be highly different from addressing a human (e.g. in terms of loudness). Also, avoiding cables and invasive attachments to the users might limit the influence of the experimental setup on the naturalness of the behavior, and the interaction might reflect patterns similar to that of a non-rigged natural interaction.
- Two high definition video cameras were used to record the setup and the interlocutors from two different angles, for future use and for annotation purposes. These cameras were used for (a) capturing tutor's behavior and (b) the entire scene.
- Six rectangular cardboard cards are designed and used as part of the game. The design of the cards was strategic in that it was made also to provide the dialogue system with context, and lower the demand for very robust speech recognition to infer the context of the game (e.g. When using vision-based card tracking, a system can infer the card under discussion, and provide spoken content related to it, without the need to understand what the users are saying).

Figure 1 shows a sketch of the physical setup to the left, and a snapshot of it with real users to the right.

[1] http://www.microsoft.com/en-us/kinectforwindows/
[2] http://www.dev-audio.com/products/microcone/

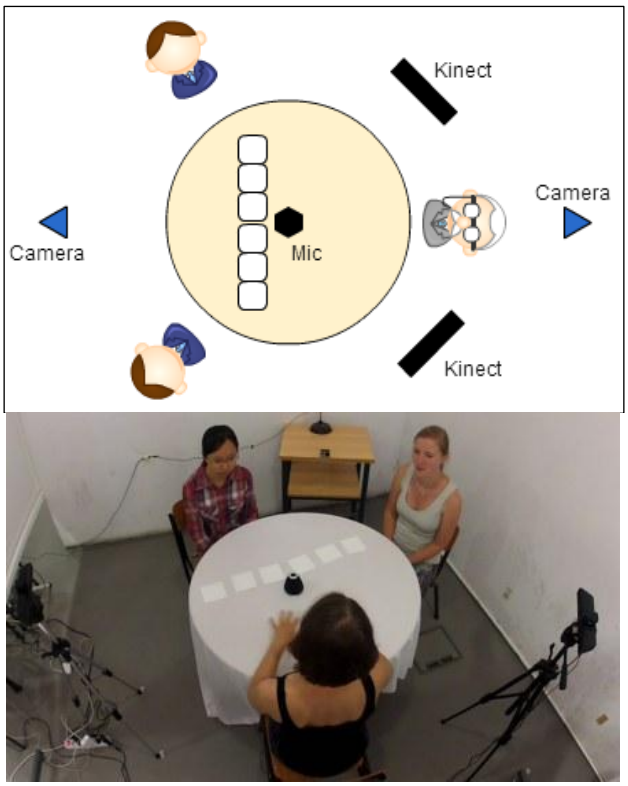

Fig. 1. Sketch (top) and a photo (bottom) of the experimental setup

Fig. 2. Six cards with pictures of objects for the task

The participants were asked to discuss each of the six cards and rank them in terms of importance. The motivation behind this task was to socially interact, exchange information, and collaboratively find a solution for the set problem. The members of each team collaborate together to solve the task. The tutor leads, coordinates, and gives comments to the team members while solving the tasks using several real-time strategies, e.g. using a *neutral* or *active* tutoring approach.

The described survival exercise was used for two reasons:

1. Groups first had to make descriptive judgments regarding the "value" of each item; and then they had to make judgments about the relative value of each item to their survival chances. Thus, both members of a group had to collaboratively participate in the conversation.
2. An important issue was the ability to compare participants' results with a right answer for the task, which was published by *the Crew Equipment Research Unit at NASA*. Group effectiveness was measured as a simple inverse function of the unit weighted sum of the absolute differences between the ranks assigned and the correct ranks. As we used a simplified version of the task, the overall performance of each group was evaluated in terms of time of task completion, in addition to the effectiveness.

4 A Corpus of Multiparty Tutoring Behavior

Recently, the research community has witnessed the birth of several large efforts towards the creation of large-scale multimodal corpora [11-15], promising that modeling verbal and visual signals in dialogue will not only advance the understanding of human-human language exchange, but also allow for the development of more intelligent and aware dialogue systems to be used by digital entities (such as ECAs and robots). However, there exist a multitude of design decisions (such as the dialogue task, the spatial setup, the captured signals) that limit the ability to easily move from human-human dialogues to human-machine dialogues. For example, dialogue tasks that heavily depend on the semantics in the speech signals (the content of the spoken interaction) will demand high requirements on speech understanding systems that can deal with conversational speech – a technology that has not yet been matured.

4.1 Users and User Setup

Before the recording session, participants were asked to complete a Big Five personality test [16]. The Big Five personality traits are five broad domains that are used to describe human personality. The Big Five factors are 1) openness to experience, 2) conscientiousness, 3) extraversion, 4) agreeableness, and 5) neuroticism. Openness reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety. Conscientiousness shows a tendency for self-discipline. Extraversion is the personality trait of seeking fulfillment from sources outside the self or in community. High scorers in extraversion section tend to be very social while low scorers prefer to work on their projects alone. Agreeableness reflects a tendency to cooperate and adjust behavior to suit others. Neuroticism is the personality trait of being emotional and refers to a degree of emotional stability. We used the personality test because research indicates that personality traits and variables like self-efficacy self-esteem, locus of control, emotional stability, extraversion, conscientiousness, positive affectivity, negative affectivity, optimism, proactive personality [17], highly impact human work

results and performance. In addition, such factors as low neuroticism in combination with high extraversion characterize work engagement [18].

For the experiment the groups were formed according to participants' personality-test results, so that one of two team members scored high on extraversion and the other one scored low. The average difference between participants on the extraversion dimension was 28 points.

The human tutor was instructed to behave in a *neutral* way with four out of eight groups. A *neutral* tutor had to deliver material in a clear and concise manner so that participants could understand what they were required to do. However, a neutral tutor didn't need to make his/her communication either interesting or enjoyable. A neutral tutor had to answer all the students' questions, coordinate their activity and explain what to do next, but a neutral tutor didn't have to try to engage students and motivate them. A neutral tutor did not need to be friendly, supportive or welcoming. For the latter four groups, the tutor was asked to behave in a way that best represents the approach of an *active* tutor. An *active* tutor had to be dedicated to a student's success, had to deliver material in an interesting manner so that students could enjoy it. An active tutor had to be supportive, friendly and welcoming and always providing a positive feedback to the student.

Eight recording sessions were performed; each session resulted approximately in 10-15 minutes conversation. Afterwards the participants were asked to fill in a Tutor Assessment Questionnaire. The assessment questionnaire was based on the User Experience Questionnaire UEQ [19] and it consists of twenty four pairs of contrasting characteristics that may apply to the tutor. The numbers between the characteristics represent gradations between the opposites. A seven-step Likert scale is used for gradation in order to reduce the well-known central tendency bias for such types of items. Please refer to Appendix A for the full questionnaire.

Twenty four characteristics were organized into groups, suggested by [19]. These groups were Attractiveness (examples for items: pleasant, enjoyable), Perspicuity (clear, easy to understand), Efficiency (fast, organized) and Dependability (supportive, meets expectations). We also had an additional group called Tutoring with items specific to the tutoring approach, e.g. motivating, holding the attention, giving feedback on the work's quality.

Validity of the used questionnaire was tested by measuring the consistence of each group, as proposed by the original UEQ [19]. The Cronbachs Alpha-Coefficient [20], for Attractiveness, Efficiency, and Perspicuity and Tutoring groups was between 0.72 and 0.95. There is no generally accepted rule on how big the value of the coefficient should be, however many authors assume that a scale should show an alpha value >0.7 to be considered as sufficiently consistent. Based on the high value of the Cronbachs Alpha-Coefficient we assume that the given groups of items in the questionnaire were consistent and that our participants in the given context interpreted the items in an expected way.

The data collected after 8 sessions (four groups with an active human tutor and four – with a neutral one) with 16 subjects an average assessment results were higher for an active tutor in all the assessment sections – attractiveness, perspicuity, efficiency, dependability and tutoring, as shown in Figure 3.

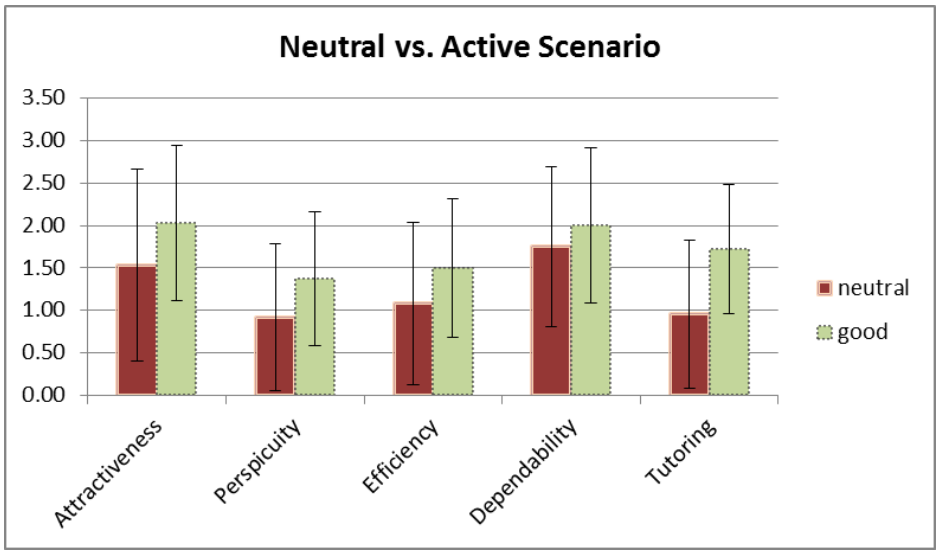

Fig. 3. Differences in tutor's assessment for an active and a neutral tutor

Although differences between active and neutral tutoring approaches didn't influence significant differences in tutor's assessment results, these two different tutoring approaches caused different overall performance for the groups of subjects with different personalities. Figure 4 shows that in case of active tutoring, when the average personality difference between group members is high, the overall task score is low, which means that the performance of that group is better. On the contrary, if the average personality difference between group members is low, the overall task score of that group is high, which means that the performance of that group was worse. In case of neutral tutoring, however, there was no such an inverse dependency between personality differences and overall task score: the more different are the group's members according to their extraversion the lower is their task performance.

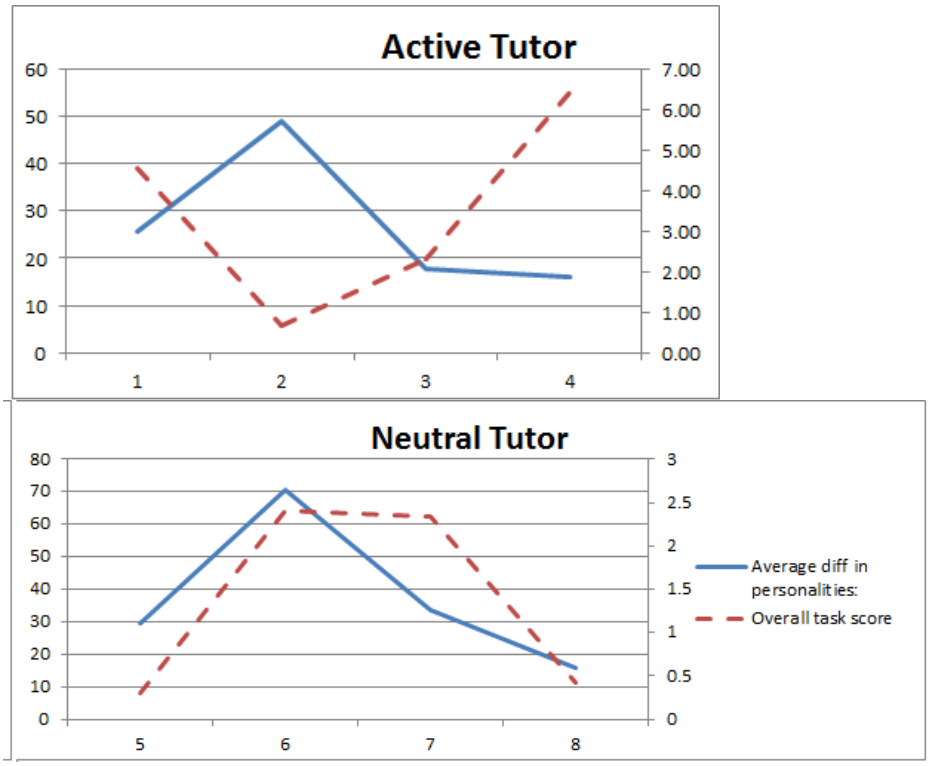

Fig. 4. The relation between differences in personalities and an overall task score in the case of active and neutral tutoring approach

Thus, according to the presented data we can argue that active tutoring eliminates differences in personalities and helps groups with higher personality gaps achieve better results in a collaborative task.

The recorded corpus was named the ***Tutorbot Corpus.***

4.2 Corpus Description

The eNTERFACE'13 ***Tutorbot*** corpus is of approximately 82 minutes overall duration and it consists of 8 sessions between a human tutor and two human participants. As described in the previous section, it includes equal samples of experimental conditions, i.e. 4 sessions of active tutoring interactional behavior and 4 of neutral. The tutor was the same subject in all sessions and was trained to express and prompt the appropriate conversational behavior according to each experimental condition. The participants, different in each session, were not informed about the task nor the goal of the experiment (participants of other projects in the eNTERFACE workshop). Depending on the availability of subjects, pairing subjects aimed at maximizing coverage in terms of personality traits (with a focus on the *extraversion* dimension) and gender. Details on the corpus are shown in Table 1.

Session	Participants (Male/Female)	Duration	Tutor scenario	Extraversion diff.	Task Score
1	M-M	13.28	Active	37	4.56
2	M-M	15.44	Active	44	0.66
3	F-F	07.08	Active	5	2.30
4	M-M	10.43	Active	21	6.45
5	M-F	07.15	Neutral	9	0.30
6	M-M	09.39	Neutral	74	2.40
7	M-M	08.07	Neutral	13	2.34
8	M-F	10.05	Neutral	21	0.42

Table 1. The Tutorbot corpus description

4.3 Annotation Process

The data collection was manually annotated with regards to the conversational behavior of the tutor. Since the goal is analyze multimodal strategies employed by the tutor to manage the conversation, the annotation was focused on describing the form and functions of the related verbal and non-verbal signals employed.

Multimodal interaction in both its two-party and multi-party dimensions is substantially related to the functions of feedback and turn management. Turn-taking mechanism has been thoroughly studied in terms of modeling the organization of turns in conversation [21], non-verbal cues such as gaze and gesture regulating turn taking in interaction [22], as well as the relationship between turn-taking and attention [23]. Multiparty turn-taking in dialog systems has been addressed with regards to the development of computational frameworks able to handle multiparty floor coordination, continuations, etc. [24]. The above studies focus on ways in which turn allocation is performed, rules that apply in transition-relevant places as well as aspects of collaborative and non-collaborative interactions such as interruptions and overlap resolution devices, both from verbal and nonverbal perspectives.

Communicative feedback gives evidence of the collaborative nature of dialogue while the participants give verbal and non-verbal signs that they follow the flow of the discussion; they perceive, understand, agree or not with the message conveyed; they might express the willingness to take the turn or give support to the speakers to go on with their turn. Feedback has been addressed in its linguistic dimension through a robust theoretical framework [25] and from a multimodal point of view with an emphasis the investigation of the effect that a combination of cues (e.g. morphological categories, prosody, gaze) might have on the production of feedback [26]. Moreover, there are attempts to describe feedback in different social activities or other contexts or model it for purposes of behavior simulation [27].

4.4 Annotation Scheme

The annotation of the recorded video sessions was performed in ELAN[3] [28]. An annotation scheme was employed to cater for all the features that need to be represented for the task at hand. The scheme is heavily based on widely-used labeling sets used for annotating multimodal interaction [29, 30], and was tailored to the needs of the task. Specifically, the goal of the annotation was to account for multimodal behavior including verbal and nonverbal signals as well as conversational structures and functions expressed in a multimodal way. Signals that have a clear communicative function are included in the annotation scheme as follows:

Speech Activity. The tutor's speech was transcribed with the goal to export utterances that the robot would use to manage the interaction. A comparison of the transcriptions was also planned, to distinguish patterns of verbal content when referring to specific subtasks in the discussion, i.e. in introducing the task, giving hints, instructing the participants to order cards, etc. as well as to discover substantial differences of verbal content in active and neutral tutor scenarios. This level includes also the transcription of verbal back-channeling (grunts such as "yeah", "ehm", "aha") the tutor may express.

Dialogue Acts. The tutor's speech activity was attributed a label of a dialogue act describing the communicative action which the tutor performs. The purpose is twofolded: (a) to identify dimensions of interaction that dialogue acts may address and (b) to functionally segment the dialogue. Since in this experimental setup the identification of the addressee is of primary importance, the information-seeking functions (i.e. questions) are categorized not in terms of question types (e.g. yes/no question, wh-question), but in terms of addressee: questions to *speaker*, *listener*, or *both participants*. A crucial part of this scenario is the cues that the tutor provides to help the participants elaborate on the cards description and their importance (i.e. *hint*). Introductory parts where the tutor asks the participant to perform an action or clarifications given throughout the discussion are labeled as *Instruction/Request*. Finally, the scheme caters for *answers* that the tutor gives to the participants or utterances of *agreement* or *disagreement* with them.

Turn Management. Values in this level describe the way the tutor regulates the interaction by taking, holding and assigning the turn. Again, the values apply to both verbal and non-verbal behavior of the tutor. Different values exist for normal transition of turns (*take*, *accept*, *complete*, *offer*), as well as for phenomena related to aberrations from the turn-taking rules, such as interruptions and overlapping talk (i.e. *grab*, *yield*, *hold*). A distinct value of *backchannel* is also included to differentiate backchannel cues from content utterances.

[3] ELAN (http://www.lat-mpi.eu/tools/elan/)

Feedback. Labels related to feedback are attributed horizontally to cover both functions of verbal and non-verbal attestations of feedback, i.e. either through back-channeling and expressing evaluations, or through head movements and facial expressions such as nodding and smiling. The set consists of labels describing whether the tutor gives or elicits continuation, perception and understanding, and whether he/she agrees or not with what the participants say.

A large part of the annotation scheme is related to the annotation of the non-verbal modalities. Since the goal of the annotation is to identify important features and patterns to be modeled in the robot, the modalities in question are restricted to descriptive and functional values of the head movements, facial expressions and facial gestures, cues that are considered of high importance to the regulation of the interaction as well as the expression of feedback. Each non-verbal signal of the ones listed below is first identified on the time axis and it is marked according to its form. Subsequently, the functions of each identified signal is marked, i.e. whether it has a feedback or a turn management purpose.

General Facial Expression. The tutor's facial expressions are indicative of his/her state of mind towards the speakers as well as of the level of perception of the discussion. *Smile* and *laugh* are employed to show agreement, encouragement and satisfaction, while *scowling* denotes doubt, disagreement or unpleasantness.

Head Movement. The form and the direction of the head movement are important for establishing feedback and turn regulating functions. For example, head *nodding* may have an acknowledgement function, by providing support to the speakers that their contribution has been perceived and that the conversation may proceed. Head *turn* is always linked with gaze to determine attention and speaker turn assignment. *Shaking* is a sign of disagreement or doubt, while *tilting* the head or moving it *forward* and *backward* may be signals reinforcing the tutor's message.

Gaze. The identification of gaze direction is of primary importance since it defines the addressee of the tutor, the goal of his/her attention and can be a clear indicator of turn assignment. The scheme distinguishes between attentive gaze of the tutor to the speakers on the *left* and on the *right* respectively. Such values may be attributed i.e. when the tutor gazes at the speaker to provide feedback, but also towards the listener in an attempt to elicit feedback or to offer the turn. Attentive gaze at the *objects* (cards) is also substantial, since it indicates the tutor follows the task process. Non-communicative gaze shifts can be labeled as *glances*.

Eyes. Variations in eye openness may indicate surprise or enthusiasm (*wide open*), as well as contemplation, interest, attention or disagreement (*semi-closed, blink*).

Eyebrows. *Raising* eyebrows is often employed to show involvement, encouragement, attention and surprise, whereas *frowning* may denote doubt, disagreement or contemplation.

Mouth. An *open* mouth is annotated as a sign that the tutor is attempting to take the turn when a participant has the floor. A *closed* mouth with protruded lips may function as a feedback signal for agreement, together with head nodding.

Cards. Finally, a dedicated layer to the id of cards that are being discussed is included in the annotation scheme, so that the boundaries of each card object are clearly identified.

A tabular representation of the scheme may be found in Appendix B, Table 4. Figure 5 also shows a snapshot of the annotation program and process showing the tutor in the video.

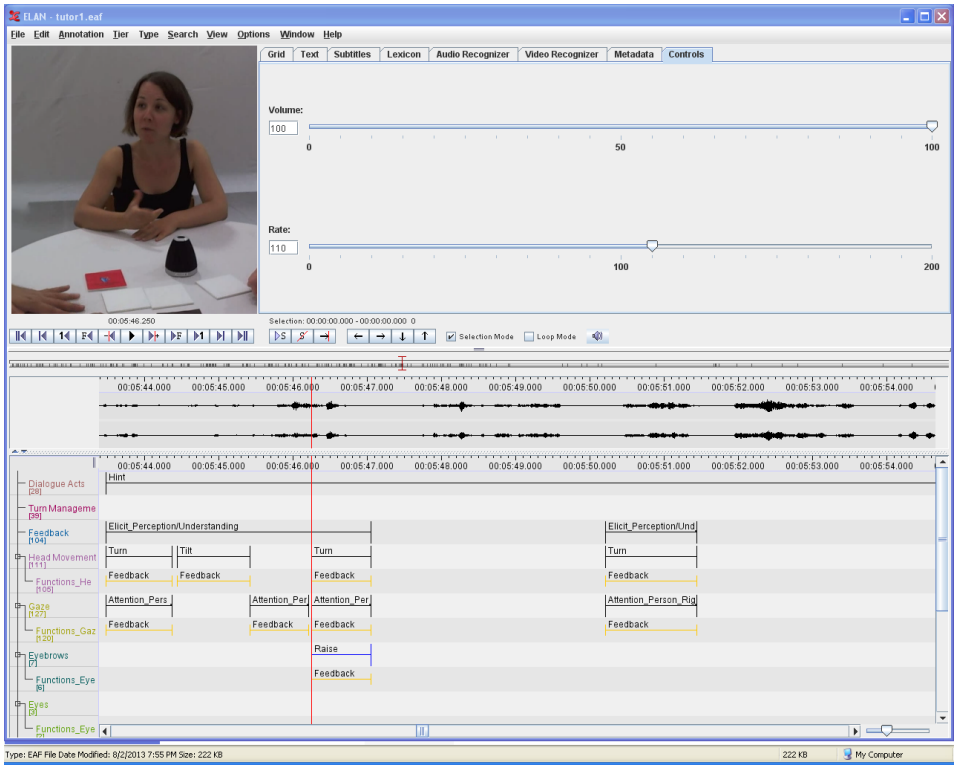

Fig. 5. A screenshot of the annotation software. The video in the frame shows the tutor, along with the manual annotations over time

4.5 Data Analysis and Conversational Management Strategies

The data collected by all devices together with the manual annotation were analyzed to model the conversational management strategies employed in both conditions of *active* and *neutral* tutor. A set of different parameters was examined to attest the interrelation of low-level signals such as voice activity, gaze, facial movements etc. and their timing with functions of turn management and feedback. Furthermore, differences in the tutor's dialogue acts and turn management behavior (in terms of frequency and different values employed) were investigated.

Our results indicate that turn management behavior conveys essential and richer information compared to the dialogue acts types used, i.e. the timing and the conversation managing action of what is said matters more than the actual content per se. For example, the number of turn offers as well as turn accepts is relatively higher in the *active* tutor condition than in the *neutral* one (42 vs. 8 and 33 vs. 13 respectively). We also hypothesized that: (a) the number of dialogue acts such as hints or instructions the tutor gives will be higher for the *active* tutor condition than for the *neutral* one and (b) the tutor will employ more turn management features in the *active* tutor condition than the *neutral* tutor condition. Concerning hypothesis a) we found a difference in the number of hints between *active* and *neutral* tutor condition (29 vs. 27 hints) and concerning hypothesis b) we also found that the number of turn grabs is higher in the *active* than in the *neutral* tutor condition (13 vs. 8 turn grabs).

Overall, an important finding derived from the corpus verifying our hypotheses with regards to the experiment design is that almost all feature cases account for the expected interactional behavior in an *active* or a *neutral* tutor. A sample of statistics calculated on features is shown in Table 2.

Feature	Active tutor	Neutral tutor
Avg. time of all conversations	11.76(3.69)min	8.77(1.3)min
Avg. time on each card	1.29 (0.66)min	0.85 (0.49) min
No. of hints in all conversations	7.25(2.06)	6.5 (1.91)
No. of agreements	2 (1.82)	2.5(1.91)
No. of disagreements	2	1
No. of instruction/request	4.25 (0.95)	3.25 (0.5)
No. of turn grabs	3.25 (2.06)	2.25(2.06)
No. of turn offers	10.5 (8.38)	4.5 (5.74)

Table 2. Statistics of features presented in the following order: *mean* (*standard deviation*)

5 Building the Embodied Tutoring agent

5.1 The Furhat Robot Head

The embodied agent used as the tutor in this project is the Furhat robot head [9]. Furhat was built to study and evaluate rich and multimodal models of situated spoken dialogue. Furhat is a robot head that consists of an animated face that is projected

using a micro projector on a three dimensional physical mask that matches in design the animated face that is projected on it. The state of the art animation models used in Furhat produce synchronized articulatory movements in correspondence to output speech [31], and allow for highly accurate and realistic control of different facial movements. The head is also supported with a 3DOF neck for the control of its head-pose.

The solution to build a talking head using the technique used in Furhat is superior in that: 1) Using a three dimensional head allows for situated and multiparty interaction that is not possible to establish accurately with avatars projected on two dimensional surfaces, thanks to its ability to eliminate the so-called Mona Lisa gaze effect – an effect that results in a loss of the orientation of 2D portrait in physical space, resulting in that a viewer of a 2D face perceives the face rotated in the same angle no matter where the viewer is standing in relation to that portrait [32,33], and 2) The use of facial animation instead of other mechatronic solutions to build robot heads enables the use of highly advanced and natural dynamics that are not so easily possible with mechanical servos and artificial skin, thanks to the advanced in facial animation techniques [31]. Furhat, in addition to being a platform to implement models of spoken human-human interaction, has become a vehicle to facilitate research on human-robot interaction, such as studying the effects of gaze movements in situated interaction [34], audio-visual intelligibility of physically three dimensional avatars [35], and effects of head-pose on accuracy of addressee selection [36]. Figure 6 shows some snapshots of the Furhat robot head.

Fig. 6. Snapshots of Furhat[4] in close-ups

5.2 The IrisTK Multimodal Multiparty Authoring Platform

To orchestrate the whole system, the IrisTK dialogue platform was used [10]. IrisTK is XML based dialogue platform that was designed for the quick prototyping and development of multimodal event-based dialogue systems. The framework is inspired by the notion of state-charts, developed in [37], and used in the UML modeling language. The state-chart model is an extension of finite-state machines (FSM), where the current state defines which effect events in the system will have. However, where-

[4] For more info on Furhat, see http://www.speech.kth.se/furhat

as events in an FSM simply trigger a transition to another state, state charts may allow events to also result in actions taking place. Another notable difference is that the state chart paradigm allows states to be hierarchically structured, which means that the system may be in several states at the same time, thus defining generic event handlers on one level and more specific event handlers in the sub-state the system is currently in. Also, the transition between states can be conditioned, depending on global and local variables, as well as event parameters. This relieves state charts from the problem of state and transition explosion that traditional FSMs typically leads to, when modeling more complex dialogue systems.

IrisTK is based on modeling the interaction of events, encoded as XML messages between different modules (a module can be a face tracker that transmits XML messages about the location of the face of a user). The design of module based systems is crucial in this system and in other multimodal dialogue tasks. Such systems are multidisciplinary in nature and researchers typically working on one of the technologies involved in such a system can be isolated from the details of the other technologies. This increases the need for the development of higher level, technology independent dialogue management that can allow for the communication of the different tools and technologies involved (Automatic Speech Recognizer - ASR, Text-To-Speech systems - TTS, Face Tracking, Facial Animation, and Source Localization). These technologies can be (and are, in this project) run on one or several machines.

IrisTK comes with several tools that support the communication of XML events in between programs, and over a network. This would allow the dialogue management to rely merely on XML events, while being able to be completely blind to the different programs that generate and consume these events. This also allows for the replacement of one or more program, or technology, without the need for any customization of the dialogue flow.

Relying on the principles of modular design and XML event communication protocol, we describe in the following the different sensory technologies that generated events, which in turn are consumed by the dialogue management flow (See Section 8 for design of the dialogue flow).

5.3 Modeling of Sensory Data

Voice Activity Detection with the Microcone™. In addition to providing audio, the Microcone™ also provides a stream of the current microphone activation status for each of its six audio channels. We used this stream to implement a voice activity detector (VAD) module for the system, by mapping microphone activation status transitions to start and end of speech.

The devised module generates a message when a subject starts speaking and when a subject stops speaking. Each message contains information about which subject triggered the message, and the amount of time passed since the previous transition, i.e. the length of silence before start of speech, or the length of the utterance at end of speech. Example XML events are presented below. Each event has a name, and a set of typed parameters. The following example shows two different events - a speech

onset and speech offset time, providing parameters on which subjects is concerned with this event.

```
<event xmlns="iristk.event" name="sense.speech.start”>
     <string name="location">Left</string>
     <float name="silence">4.5</float>
</event>
<event xmlns="iristk.event" name="sense.speech.end>
     <string name="location">Right</string>
     <float name="length">2.25</float>
</event>
```

As the experiment setup has the participants in fixed locations, the identity of a speaker can be mapped to an audio channel. The module keeps track of the current activation state of each channel in use, and creates events on activation state transitions, provided the state has been stable for a tunable time period. The tunable threshold allows for brief moments of silence in continuous utterances, and prevents short isolated sounds from triggering start of speech detection.

Visual Tracking. The behavior of the subjects was tracked partly using Kinect sensors (“Kinect for Windows”). In the experiment setup, one sensor was used for each subject and each sensor was placed in front of the subject it was monitoring. The intention was to have the subject facing the sensor, keeping the expected head pitch and yaw angles between +/-30 degrees.

The physical locations and orientations of the subjects’ heads, as well as parameters describing the facial expressions, were tracked using the Microsoft Face Tracking SDK (“Face Tracking”). Data from the head tracking allowed the system to have Furhat to direct its gaze at subjects, and to estimate the visual attention of the subjects. In addition to the tracking of heads, the poses of the subjects’ upper bodies were tracked as well, using the Kinect for Windows SDK skeleton tracking in seated mode (“Tracking Modes”). Skeleton data with ten tracked upper-body joints for each subject was collected for future use.

```
<event xmlns="iristk.event" name="sense.head">
    <string name="sensor">kinect_left</string>
    <string name="agent">Left</string>
    <float name="position.x">1.2</float>
    <float name="position.y">2.345</float>
    <float name="position.z">3.45678</float>
    <float name="rotation.x">1.2</float>
    <float name="rotation.y">2.345</float>
    <float name="rotation.z">3.45678</float>
    <float name="au.lipstretcher">-0.2</float>
</event>
```

Visual Attention Estimation with Kinects. We wanted to provide the system with information about the subjects' visual focus of attention, which can be inferred from gaze direction. Complete gaze direction is, however, not available in the case of our setup, so an alternate method is required. One way of estimating the visual focus of attention without gaze is to use head pose information as a surrogate. This alternative was explored in [38], who in a round-table meeting scenario with four participants show an average accuracy of 88.7% for the estimation of focus of attention from head orientation alone. The contribution of the head orientation to the overall gaze was on an average 68.9%. A study in [39], expanded the meeting scenario with additional targets for visual attention, and found that the different targets for visual attention need to be well separated in order to achieve good estimation performance. In our experiment setup the potential targets for a subject's attention is the tutor, the other subject, or the card on the table. The low target count, combined with the constraints of the experiment setup, suggest that we can use head poses as a good estimate for subjects' visual focus of attention in our experiment, similar to a setup in [40].

The devised visual attention module for our setup accepts head tracking data from the Kinect sensors and generates one message whenever the estimated visual attention target of a subject has changed. The message, illustrated below, contains information about which subject the message refers to, the direction of the subject's head and the estimated target of the subject's visual attention. Before the target is considered changed and a message is sent, the estimated target must be stable for a tunable period of time.

```
<event xmlns="iristk.event" name="sense.attention">
    <string name="agent">Left</string>
    <string name="direction">left</string>
    <string name="target">Right</string>
</event>
```

Since the physical setup ensures that each Kinect sensor is covering exactly one subject, the identity of detected heads could be derived from the sensor detecting it. The target of visual attention of a subject was estimated based on the pan and tilt of the subject's head. Each possible target was specified as a region defined by minimum and maximum angles to the target relative to the Kinect sensor located in front of the subject. The region boundaries for visual attention targets were calibrated manually for this experiment, a method we believe to be rather reliable for the current purpose due to the well-structured physical setup. We do, however, intend to improve the boundary definition by using more data driven methods for clustering in the future. Combining the microphone array and the visual attention classifier, the system can be informed about the speaker and the addressee at any given point in time. Figure 7 shows a visualization of the system in action, showing Speaker Left speaking to Subject Right, while subject right is looking at the tutor.

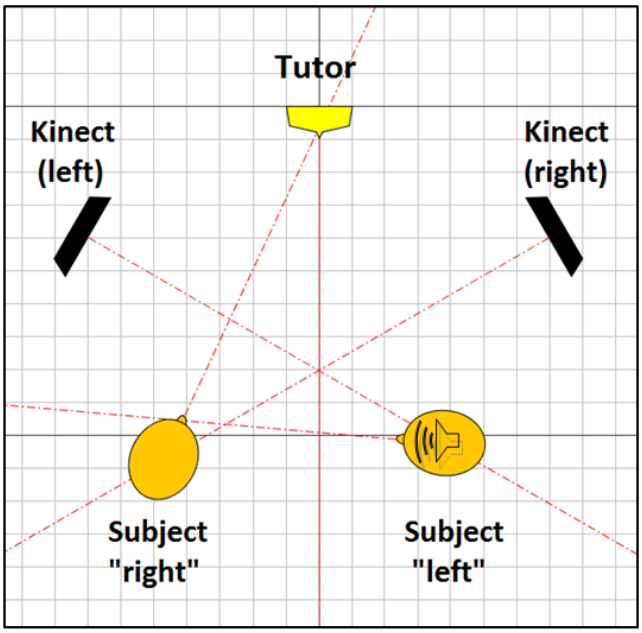

Fig. 7. A visualization of the perceived activity of the situated interaction. The figure shows the tracked horizontal head rotation (pan) of each subject, and the voice activity detection (highlighted by an image of a speaker). The figure shows that the subject (left) is currently speaking.

Acoustical Prominence Detection. In addition to the Speech Activity, and the Visual Attention modules, a prosodic analysis module was developed to estimate important segments in the users input using their prosodic features. Information about prosodic prominence can in principle enhance the speech recognizer by conditioning the syntactic parsing. In addition to that, it can also give important information to the agent, which in turn can show more contextually aware gestures in relevance to the users' prosodic contours [41]. In this work, we wanted Furhat to generate nonverbal gestures (as eyebrows raises) in response to prominent segments in the users speech, in order to show attentive behavior (such strategies have been shown to be functional in active listening experiments, c.f. [42]).

Acoustical prominence is perceived when a syllable or a word is emphasized so that it is perceptually salient [43]. Detecting the acoustical prominences can be very useful in Human-Robot interaction (HRI) scenarios. For example, the salience of the emphasized words uttered by human beings can be used as cues for triggering feedback signals generated by robots. The feedback signals can be acoustical by saying yeah or mmm, visual by raising eyebrow or smiling, or multimodal [44]. These feedback signals make the interaction between humans and robots more natural and increase humans' engagement in the conversation. Moreover, in multiparty dialogues, the frequency of the detected prominences can be used as a reasonable feature for detecting and balancing the conversational dominance of the dialogue participants.

In order to automatically detect acoustical prominences, some low level acoustic prosodic features should be first extracted. Features like F0, energy, and duration have shown a good success in automatic prominence detection [43]. Mapping these extracted prosodic features to the acoustical prominence, which is mostly defined based on linguistic and phonetic units, is conventionally done using annotated databases and supervised machine learning algorithms like neural networks (NN) [43], and hidden Markov models (HMM) [45]. In this work, however, we have used a modified version of the unsupervised statistical method applied in [46]. The main idea of this method is developed based on the prominence definition introduced in [47].

In [47], prominence is defined as the speech segment (syllable or word) that stands out of its environment. In order to realize this definition, we define a relatively short moving window in which the current speech segment (e.g. syllable) lies and another longer window for its preceding environment. We can simply detect the salience of the current speech segment by calculating a discrimination distance between the prosodic features that lie in it and those located in its preceding (environmental) window. Prominence is then detected if the distance is larger than a pre-defined threshold, which means that the current segment is salient and stands out of its environment.

The discrimination distance can be deterministic such as the Euclidean distance between the prosodic features' mean vector of the current and the environmental window, or probabilistic to take into account the uncertainty (covariance) of the feature vectors. A good candidate for the probabilistic discrimination distance is the Kullback-Leibler (KL) divergence [48]. By assuming that the prosodic feature vectors in the current local and the past global window are modeled by Gaussian distributions, the KL divergence can be computed via [49]:

$$D_{KL}(\mathrm{N0}||\mathrm{N1}) = \frac{1}{2}\left(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0)(\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - k - \log\left(\frac{\det(\Sigma_0)}{\det(\Sigma_1)}\right)\right) \quad (1)$$

where μ_0, μ_1, Σ_0 and Σ_1 are the mean feature vectors and the covariance matrices of the current local window and the past global window, respectively. In (1.1), k is the feature vector dimension.

However, one problem of applying the KL divergence here is the difference in the estimation reliability of the global and the local window parameters. This difference arises due to the convention of choosing the local window length shorter than that of the global window. For that reason and the fact that the KL-divergence is non-symmetric, we have used instead a modified version of the T^2 Hotteling distance:

$$D_H(\mathrm{N0}||\mathrm{N1}) = \frac{L_0 L_1}{L_0+L_1}\left((\mu_1-\mu_0)^T W\, \Sigma_{0\cup1}^{-1}(\mu_1-\mu_0)\right), \quad (2)$$

where L_0 and L_1 are the length of the local and the global window and $\Sigma_{0\cup1}$ is the covariance matrix of the union of the samples of the local and the global windows. The main function of the added weight matrix W here is to give prosodic features different importance. However, if all the prosodic features are of the same importance

then the weight matrix W will be the identity matrix I and the T^2 Hotteling distance in (1.2) reduces to its standard form in [50].

Implementation aspects.
The prosodic features used in this experiment are the fundamental frequency F0 and the energy E. To extract these features, we have implemented a real-time multi-channel prosodic feature extractor module. This module extracts the short time energy in dB via

$$E(t) = 10\log\left(\frac{1}{K}\sum_{k=0}^{K} x^2(k,t)\right) \tag{3}$$

where the $x(k,t)$ is the k^{th} sample of the tth frame and K is the frame length. The F0 in this module is extracted according to the pitch tracking algorithm YIN [51] (e.g. see Figure 8). The prosodic features are extracted from short time frames of length 50 ms with 50% overlap between consecutive frames. In order to compensate the outliers, the output of the pitch tracker is applied to a median filter of length three.

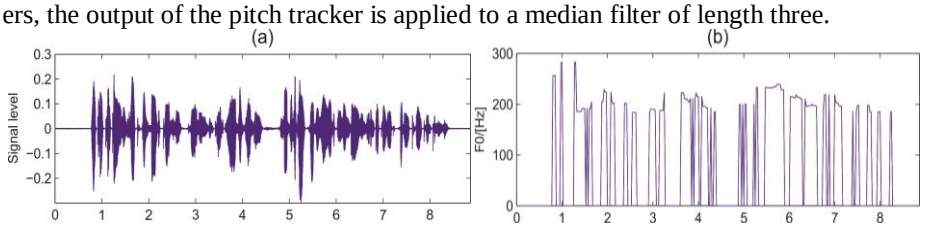

Fig. 8. (a) Exemplary input signal to the real time F0 tracker. (b) The estimated F0 using the YIN algorithm.

In [52], it has been shown that the average duration of vowels in heavily stressed syllables is between 126 and 172 ms. Thus, the length of the local window has been tuned in this range so that the performance of the prominence detector is optimized. The length of the global window that models the environment of the current acoustical event is chosen to be seven times larger than the length of the local window length. The feature vectors used to calculate the T^2Hotteling distance in (2) are the feature vectors extracted only from voiced speech.

Visual Tracking of Game Cards. To allow the dialogue system to infer the status of the game, without the dependency on interpreting spoken content from the users, the game design employed six cardboard cards on each one of which an object was shown. Since the design of the game and the setup was not mandated by technical limitations, this allowed for flexibility to design and color the cards to maximize the accuracy of a card tracking system that is not sensitive to lighting changes. The design of the table and the game was also flexible, thus the game was designed so that subjects would place the cards in certain dedicated spots and with a certain orientation.

Detection, recognition and tracking of an arbitrary object in video stream are inherently difficult tasks. Most of the problems stem from the fact that light conditions change over time. Furthermore, abrupt motion, changes in shape and appearance and occlusions make the tasks more challenging. There is tremendous amount of research targeted at tackling these problems and many algorithms have been proposed in the literature [53].

The main requirement for the developed system is to have real-time response. There are computationally feasible methods, which can compete with the more complex ones, given a set of assumptions [53]. Since we can design the game and the flow, we can safely assume that light condition will not significantly change during the game and the shape and the appearance of the tracked objects is constant.

The dialogue system was conditioned by input from the card tracking about the timing and an identity of a new card (whenever users flipped a new card to discuss it). Another requirement from the card tracking system was that whenever the users flip all the cards and put them in order, the card tracking system needs to inform the dialogue system that a new order has been established (the tracking system should also inform the dialogue system whenever a new order is in place – this could happen if the users discuss further the cards and change their agreement).

The cards are designed so that each one has a distinct color. This enables the system to differentiate the cards by comparing their color histograms. This type of comparison is convenient because the color histogram does not change significantly with translation and rotation.

The card tracking system accepted a video stream from a video camera that is directed towards the table (Figure 9). The system allowed the experiment conductor to initialize the templates of the cards whenever needed. This is assumed to be important by the beginning of each interaction, as lighting changes might happen over large periods of time. In order to initialize the system, the user is required to define two regions of interest in the video. The first one is the region at the bottom of the table where the card under discussion resides (users were asked by the tutor agent to flip open a card and place it in the dedicated spot before they start discussing it). The second is the region in the middle of the table where all cards reside – this region of interest is used to track the order of all the cards whenever the cards are all flipped open at the end of the discussion. Figure 9 shows an image from the video stream of the camera used for the card tracking system, and illustrates a selection of the active card region of interest.

Fig. 9. Active card ROI

In order to recognize the card at the bottom of the table, first a contour detection is performed on the whole region of interest. All contours are then approximated with polygons and then each polygon is described by its bounding box. After filtering the resulting bounding boxes through predefined threshold (removing small false detections) the smallest bounding box is selected as the target object. The histogram of the crop of the target is calculated and the correlation between the target histogram and all template histograms is calculated. The closest template is chosen as the recognition result. If the recognition does not change for a predefined number of frames, the system broadcasts this decision to the dialogue manager.

After all the cards are discussed, the participants need to agree on order of relevance. This is done following the same algorithm used to track an active card. The cards are then sorted with respect to the top-left corner coordinates of their bounding box. Thus, we obtain the relevance information for each of them (e.g. the left most being the most important). Figure 10 illustrates real-time tracking and the pipeline of the implemented system.

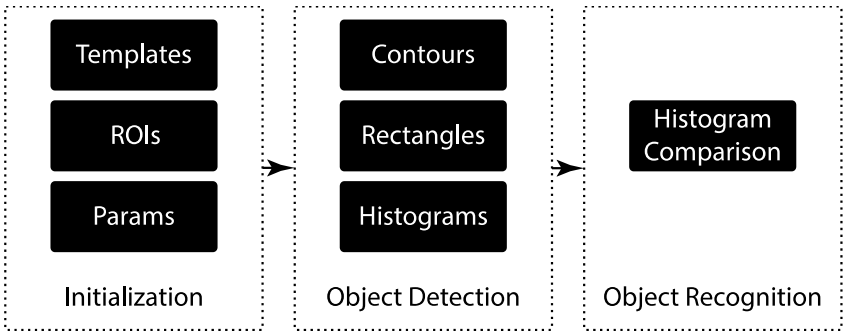

Fig. 10. Card tracking pipeline

6 Dialogue System

In this project, the IrisTK framework [10] for multimodal spoken dialog system is used to author the dialogue system, and in turn to control the Furhat robot head. This framework allows the incorporation of a series of modules and facilitates a standardized event-based communication between them. These events can represent input data (Sense), something that the system should do (Action) or feedback-loop sensation about a certain action to infer the physical state of the system (Monitor). The system architecture was designed to accommodate different setups according to the needs.

The sensory modules collect information from the environment:

- The card-tracking module is responsible for two different tasks: track the card under discussion and, when the discussion ends, provides the system with the card order. The details of how the card tracking is performed are detailed in the previous section.
- The two Kinects which track the orientation of the heads in order to infer visual attention. With the head position, the system can be informed about the location of the speakers (which would help the system establish mutual gaze when needed), and with the orientation, the system can track where the speakers are focusing their attention, and who they are addressing (for details about the visual attention module, please refer to Section 5.3)
- Finally, to combine the visual attention with the verbal activity, the setup employs one microphone array (Microcone™), composed of six microphones positioned around a circle covering in total 360 degrees. The microphones are used to perform Voice Activity Detection (VAD) for each of the participants in the dialogue.

The actuating modules perform the visible actions. The Furhat module is responsible for managing the agent's face, gaze, gestures and speech synthesis tasks. The Neck module performs head movements mainly by directing attention to the speakers, using input from the Kinect face tracking modules. Figure 11 shows a flow chart of the main modules of the dialogue system.

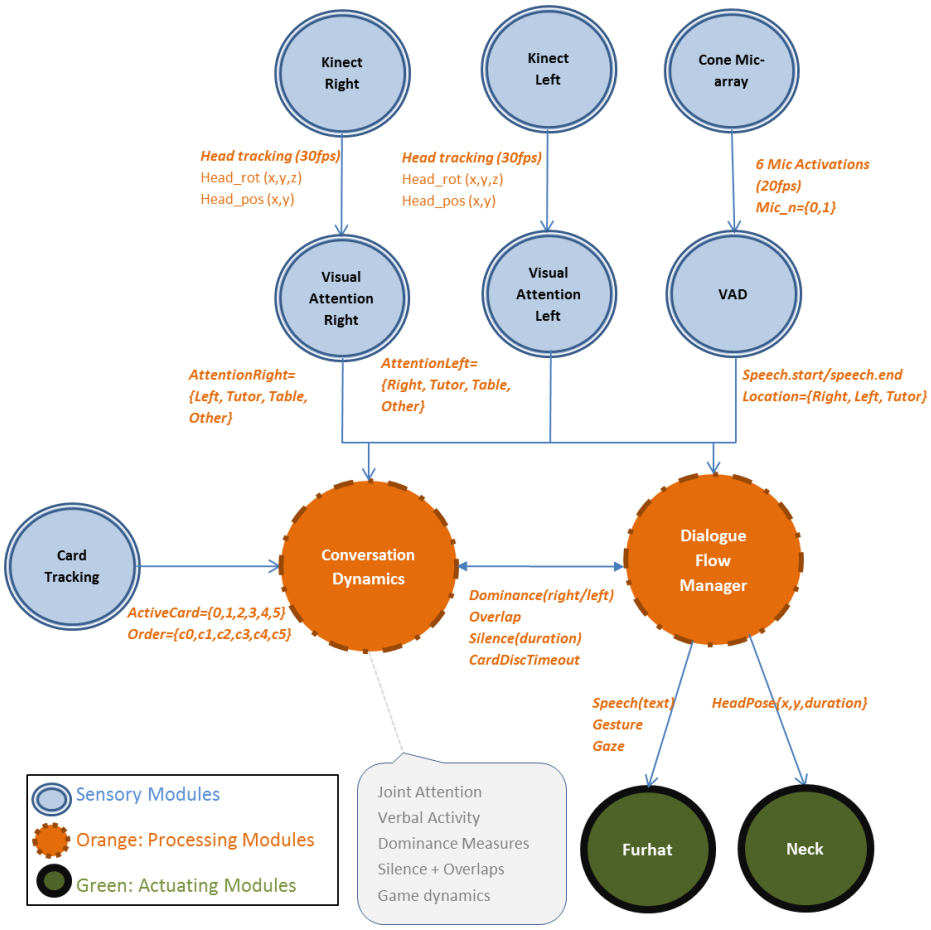

Fig. 11. An overall chart view, showing the flow of information in the system. Each circle represents an independent Module that communicates with other modules using events encoded as XML messages. Sensory modules are mainly responsible for providing input events to the system. Processing modules are responsible for modeling the dialogue using sensory events and internal state definitions. Actuating modules are responsible for activating the robot head.

6.1 Conversation Dynamics

Conversation Dynamics is a key module for Furhat's tutoring task since it computes figures in real time (updated every 1ms) that relate to conversational properties of the interaction. Such events can be looked at as the main drive that will decide when the tutor should intervene in the dialogue. The principle behind supporting the dialogue manager with a "conversational dynamics" module is to build an up-to-date model of the interaction. Such model allows the dialogue manager access to high level states of the interaction, instead of calculating them as part of its dialogue state design. This will remove the need on the dialogue system to contain dedicated States for each and

every possible combination of sensory input and context. Since the task of the system is to infer higher level conversational parameters and act on them (such as dominance, low levels of engagement), the conversational dynamics is responsible for containing variables about the verbal activity of each participant, their current visual target, and other long-term parameters, such as the percentage of silence a certain user has been in since the beginning of the dialogue. This allows the dialogue manager to access these parameters on demand. The conversational dynamics module is also responsible for firing events related to the interaction between users, for example, if users are silent for more than a specific threshold, the conversational dynamics module can send an event called "low engagement", which the dialogue flow in turn can respond to by taking the initiative and directing a question to one of the participants.

- Verbal activity is computed for both speakers. For each of them a vector with the frames where each of them spoke in the last 200ms is computed, based on the Voice Activity Detection performed by the microphone array. These vectors are used to compute the dominance that relies upon the difference in the verbal activity between the two speakers over a longer period of time. Measuring dominance is a research topic by itself and has received considerable attention, where most work has targeted the offline annotation of meeting corpora. In [54] for example, Support Vector Machines (SVMs) were used for a posteriori classification of dominance in meetings. In our scenario, we needed a real-time dominance classifier. The solution adopted was a rule-based decision, using a threshold on the difference of verbal activity between the two speakers inspired by the analysis of the recorded corpus, using the timing of tutor interruption and turn management as thresholds. If the difference in the verbal activity is above that threshold, a dominance event will be generated. The verbal activity values are reset once the card under discussion changes.
- Conversation dynamics also computes the silence information for each of the speakers separately and joint silences for all participants (including Furhat). The period since speakers started speaking is also computed for each of them. The combination of speaking times results in the overlap speech period. Both are computed since the conversation started and since the card under discussion changed.
- This module also tracks the duration of the current card discussion and total discussion times. If these reach the thresholds set, events are generated to make the system suggest a change in the card under discussion or, in case of ordering, to suggest the end of the discussion.

6.2 Flow Description

The dialogue manager is specified using a state chart-based framework defining the flow of the interaction (IrisTK flow [10]). In our experiment two different flows were created, one for the *Neutral* tutor and another for the *Active* tutor. These tutors try to map the characteristics revealed by the different tutor behaviors in the sessions with the human tutor. The complete diagram flow is shown is Figure 12. The difference between the tutors is not the flow itself, but the way the states are implemented. The first state defined in the flow is the Game state, a general state. All the other states in the flow will extend the specifications of this state, which means that the behaviors specified within this state would be available in every state that extends this one. These behaviors correspond to actions that the head must perform. The following bullets explain the structure of the dialogue system in terms of the states it occupies over time, and in relevance to the flow of the dialogue.

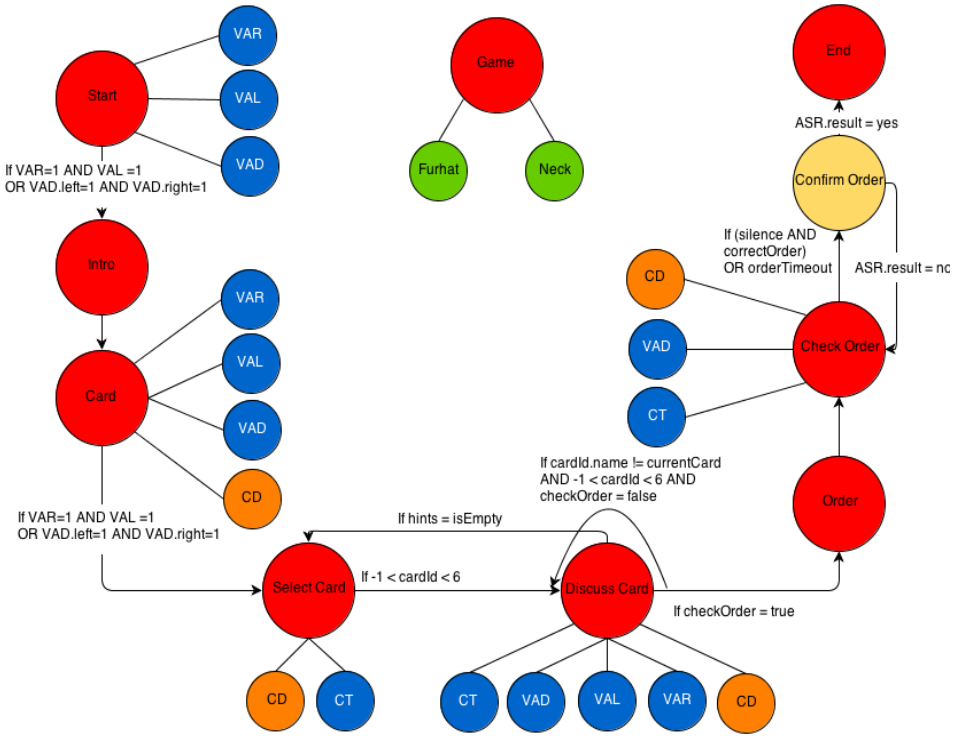

Fig. 12. Flow chart of the dialogue system states. VAD: Voice Activity Detection. VAL: Visual Attention of Left speaker. VAR: Visual Attention of Right speaker. CD: Conversational Dynamics. CT: Card Tracking input.

- The “Start” state is the initial state in the dialogue. In this state, the system greets the users for the first time and waits until both of them are detected to move on to the next state. This detection is performed either using visual cues (collected from Kinect) or audio cues (collected from the microphone array). If only one of the users is detected, the system informs her/him that they should wait for the other user to be detected in order start the discussion. Once both users are detected the flow goes to “Intro” state.
- In the “Intro” state, an explanation of the moon survival task is given to the user. After this explanation, the dialog proceeds to the “Card” state.
- The “Card” is just a step to check that the users are ready to play the game. The system checks that they are ready by detecting them using voice activity detection and visual attention state of both of them, and then moves the flow to the “Select Card” state. If they remain silent above the threshold silence time, than the system prompts a sentence to make the users speak and move to the “Select Card”.
- The “Select Card” state waits for a sense.card event generated from the card tracking module. The event contains the card id of the identified object. Until the id is valid or the users keep the silence the system is going to push them to select a valid card.
- If the valid id is detected the “Discuss Card” state is activated. In this state, the system manages the card discussion. The discussion management is different between the two types of tutor. In both of them silent periods are tracked.
- If the users have been silent for a specific time (a threshold is reached), the system is going to evaluate the user’s attention. In the human tutor dialogs, if both users were looking to the tutor and one of them asked something to the system, they are both waiting for an answer. The tutor should address both speakers and use one of the hints transcribed for this type of behavior. The hints for the objects are loaded as a stack. Once the hint is used it is popped out of the stack. It might occur that the hints for the object under discussion were all popped from the stack. In that case the system will encourage the users to pick another card.
- If only one of the users is looking at the tutor, and she/he was the last to talk, the tutor only addressed this user when answering. The same behavior was implemented, having the tutor answering towards the last user who talked and using one of the hints transcribed in the corpus for the object under discussion.
- The system is also measuring the total silence time. If the threshold is reached, the system should use one of the prompts that were used by the human tutor whenever there were long silences. These prompts should encourage the users to continue the conversation.
- There is also a timeout and minimum time for a card discussion. When the timeout is reached the system suggests the users to flip over a new card. If card event (change of card) is detected before the minimum discussion time is reached the system gives a hint about the card that was being discussed before the card event was detected, in order to make the users continue the argument about the card that they have decided to change. These thresholds were set based on the data collected in the human tutor corpus. Thus different thresholds were used for the Neutral and Active tutor flows.

- Another difference between the two configurations is how they deal with a dominance event generated by the Conversation Analysis. The Active tutor grabs the turn from the dominance speaker whenever the Conversation Dynamics module has generated the dominance event, whereas the Neutral tutor does not interrupt the discussion when there is a dominance event. An example scenario would be that one of the speakers has been speaking continuously reaching a threshold, without any interruption by the other participant. Whenever this turn length reaches a set threshold (estimated from the human recording data), the tutor interrupts the speaker by saying something like "But what do you think", or "have you though that the moon does not have a magnetic field", while turning the head towards the silent participant.
- Once all the cards are individually discussed, the Card Tracking module detects a final order of the card and generates a "cards ordering event", moving the flow to the Order state. This state simply informs the speakers that they should start ordering the cards and moves to the Check Order state. When the system is in this state, silence periods are measured. If they reach the threshold, the system verifies if the order is correct. If the order is correct the system goes to the Confirm Order state and explicitly confirms if the final order has been reached and if both participants agree on the order the game ends. The Confirm Order state is the only state in the dialog that does not extend the behaviors of the Game state. The confirmation is made using speech recognition with a simple yes/no grammar. If the order is not correct, the system gives a transcribed hint that the human tutor used in this context. In this state, when the order timeout is reached, the system recaps the current order and goes to the Confirm Order state to explicitly confirm if that is the final order. This system will repeat these steps until the users agree on the order.
- Similar to discussing the cards, the Active tutor performs the behaviors implemented in the Neutral tutor and above described, with two new features. Dominance is detected as described for the Discussion state and there is also a minimum ordering time threshold, that is, the tutor encourages the speakers to continue the ordering discussion if their discussion time is too short.

7 Discussion and Future Work

In this work, we presented a novel experimental setup, and corpus collection, and the design details of a complex multiparty dialogue system. One of the main criteria that dictated the design of the system is to keep the experimental setup as natural as possible, in order to allow users to employ natural behaviors common in human-human conversations. We also chose a task that would give the system a special role rather than trying to simulate a task-independent human-human multiparty dialogue. The choice we took in this project is to design a task that enforces certain restrictions on the interaction in a way that would give benefit to the technology employed rather than limit the interaction. The tutoring setup was set in a way to allow the tutor to give any type of information, or to choose to be passively monitoring the interaction,

lowering the expectations of the users on the "apparent intelligence" of the tutor. The design of the task also employed the use of physical objects that could be tracked reliably using computer vision. This would produce solid pieces of information regarding the content of the dialogue, allowing the system to produce context-specific and information-rich content (such as giving hints about a specific card, whenever users are silent for a specific period of time).

The design of the dialogue system is also novel in several aspects. The system can handle two users at the same time, and take their visual and verbal activity into account. The choice of building a Conversational Dynamics component of such interactions, we believe, is valid for a large set of face-to-face multiparty human-machine dialogues. Such setups target the development of human-like behaviors that will need to depend on large and long-term contexts and variables (such as dominance, involvement and engagement, etc.). Such modeling of high level conversational variables cannot be a simple extension of dialogue states.

From pilot tests with users, the system shows great potential. The ability of the system in knowing the addressee of a certain utterance produced by a user enhances the interaction significantly. We intend to carry a large user-study to evaluate the system thoroughly, in regards to conversational management strategies, using the Active and Neutral tutoring patterns and actions found in the corpus, which practically will tune the thresholds for different regulatory actions done by the robot.

The work established in this project can be regarded as a research platform to explore the effects of different conversational strategies on users. One can, for example, control certain parameters in Furhat's behavior, and tune them systematically to study large effects on the conversation, in an unprecedented way. The system for example, can attempt to bond with one user over the other by control of agreement, facial expressions, verbal and nonverbal feedback, and support with hints. The platform can also be used to study verbal and nonverbal alignment and entrainment in multiparty dialogue, where certain parameters (such as loudness, pitch, emotions, speech rate) can be controlled, and manipulated differently for each user.

The area of face-to-face multiparty dialogue is highly rich and unexplored, compared to its dyadic-dialogue counterparts. We think of this work as an attempt to design an experimental setup where different behaviors in face-to-face socially aware multiparty conversations can be studied.

Acknowledgements. This project was carried out at as part of the eNTERFACE'13 Multimodal Interfaces Workshop in Lisbon, Portugal. The project members are thankful to the organizers for providing the space and resources. Samer Al Moubayed was partly funded by the KTH Strategic Research Area - Multimodal Embodied Communication. Jekaterina Novikova is funded by the University of Bath and her participation was partly funded by the Society for the Study of Artificial Intelligence and Simulation of Behaviour. The authors would also like to thank the reviewers for their constructive comments, and the eNTERFACE participants for taking part in the data recordings.

References

1. Cassell, J.: Embodied conversational agents, Cambridge, MIT Press (2009)
2. Rudnicky, A.: Multimodal dialogue systems. In W. Minker et al. (eds.) Spoken Multimodal Human-Computer Dialogue in Mobile Environments, volume 28 of Text, Speech and Language Technology, pages 3–11. Springer (2005)
3. Clifford, N., Steuer, J. & Tauber, E.: Computers are social actors. CHI '94: Proc. of the SIGCHI conference on Human factors in computing systems, ACM Press, pp. 72–78 (1994)
4. Cohen, P.: The role of natural language in a multimodal interface. In proc. of User Interface Software Technology (UIST '92) Conference, Academic Press, Monterey, CA, pp. 143–149 (1992)
5. Cohen, P. & Oviatt, S.: The role of voice input for human-machine communication. Proceedings of the National Academy of Sciences, 1995. 92(22): p. 9921-9927 (1995)
6. Castellano, G., Paiva, A., Kappas, A., Aylett, R., Hastie, H, Barendregt, W., Nabais, F., Bull, S.: Towards empathic virtual and robotic tutors. In: Artificial Intelligence in Education, pp. 733-736, Springer Berlin Heidelberg (2013)
7. F. Iacobelli F., Cassell, J.: Ethnic Identity and Engagement in Embodied Conversational Agents. In: Pelachaud, C., Martin, J.-C., Andre, E., Chollet, G., Karpouzis, K., Pele, D. (eds.) Conf. on Intelligent Virtual Agents (IVA 2007), pp. 57–63. Springer (2007)
8. Robins, B., Dautenhahn, K., te Boekhorst, R., Billard, A.: Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? Universal Access in the Information Society (UAIS) (2005)
9. Al Moubayed, S., Beskow, J., Skantze, G., Granström, B.: Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. In Esposito, A., Esposito, A., Vinciarelli, A., Hoffmann, R., Müller, V. (eds.), Cognitive Behavioural Systems. LNCS, vol. 7403, pp. 114–130, Springer (2012)
10. Skantze, G., Al Moubayed, S.: IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In: ICMI 2012, Santa Monica, CA (2012)
11. Oertel, C., Cummins, F., Edlund, J., Wagner, P., & Campbell, N.: D64: a corpus of richly recorded conversational interaction. Journal of Multimodal User Interfaces (2012)
12. Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D.: Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., & Tapias, D. (Eds.), Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10) (pp. 2992 - 2995). Valetta, Malta (2010)
13. Al Moubayed, S., Edlund, J., Gustafson, J.: Analysis of gaze and speech patterns in three-party quiz game interaction. In: Interspeech 2013. Lyon, France (2013)
14. Paggio, P., Allwood, J., Ahlsen, E. Jokinen, K., Navarretta, C.: The NOMCO multimodal Nordic resource - goals and characteristics. Proceedings of the Language Resources and Evaluation Conference (LREC-2010). Valetta, Malta (2010)
15. Carletta, J.: "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus." Language Resources and Evaluation 41.2 181-190 (2007)
16. Digman, J.M.: "Personality structure: Emergence of the five-factor model". Annual Review of Psychology 41: 417–440 (1990).
17. Bateman, T.S., Crant, J.M: The proactive component of organizational behavior: A measure and correlates, Journal of Organizational Behavior 14 (2) 103-118 (1993)
18. Langelaan, S., Bakker, A., Van Doornen, L., Schaufeli, W.: Burnout and work engagement: Do individual differences make a difference?, Personality and Individual Differences, 40 (3), 521-532 (2006)

19. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: HCI and Usability for Education and Work, pp. 63-76, Springer Berlin Heidelberg (2008)
20. Cronbach, L.J.: Coefficient alpha and the internal consistency of tests. Psychometrika 16, pp. 297–334, (1951)
21. Sacks, H., Schegloff. E., Jefferson, G.: A simplest systematics for the organization of turn-taking in conversation. Language 50, 696-735 (1974)
22. Duncan, S.: Some Signals and Rules for Taking Speaking Turns in Conversation. Journal of Personality and Social Psychology 23, 283-292 (1972)
23. Goodwin, C.: Restarts, pauses and the achievement of mutual gaze at turn-beginning. Sociological Inquiry, 50(3-4), 272-302 (1980).
24. Bohus, D., Horvitz, E.: Facilitating multiparty dialog with gaze, gesture, and speech. In: ICMI'10, Beijing, China (2010)
25. Allwood, J., Nivre, J., Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. Journal of Semantics, 9(1), 1-29 (1993)
26. Koutsombogera, M., Papageorgiou, H.: 'Linguistic and Non-verbal Cues for the Induction of Silent Feedback', in A. Esposito et al. (eds.) Development of Multimodal Interfaces: Active Listening and Synchrony, LNCS, vol. 5967, pp. 327-336 (2010)
27. Allwood, J., Kopp, S., Grammer, K., Ahlsén, E., Oberzaucher, E., Koppensteiner, M.: The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. Journal on Language Resources and Evaluation, 41(3-4), 255-272 (2007a)
28. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: ELAN: a professional framework for multimodality research. In: 5th International Conference on Language Resources and Evaluation (LREC 2006), 1556-1559 (2006)
29. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena. Multimodal Corpora for Modeling Human Multimodal Behaviour. Journal on Language Resources and Evaluation, 41(3-4), 273-287 (2007b)
30. Bunt, H. Alexandersson, J., Carletta, J., Choe, J.-W., Fang, A.C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., Traum, D.R.: Towards an ISO Standard for Dialogue Act Annotation. In: Seventh International Conference on Language Resources and Evaluation (LREC 2010) (2010)
31. Beskow, J.: "Rule-based visual speech synthesis," in Proc of the Fourth European Conference on Speech Communication and Technology, (1995).
32. Al Moubayed, S., Edlund, J., Beskow, J.: Taming Mona Lisa: communicating gaze faithfully in 2D and 3D facial projections. ACM Transactions on Interactive Intelligent Systems, 1(2), 25 (2012)
33. Al Moubayed, S., Skantze, G.: Turn-taking Control Using Gaze in Multiparty Human-Computer Dialogue: Effects of 2D and 3D Displays. In: AVSP 2011, Florence, Italy (2011)
34. Al Moubayed, S., Skantze, G.: Perception of Gaze Direction for Situated Interaction. In: 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, The 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA (2012)
35. Al Moubayed, S., Skantze, G., Beskow, J.: Lip-reading Furhat: Audio Visual Intelligibility of a Back Projected Animated Face. In: 10th International Conference on Intelligent Virtual Agents (IVA 2012), Santa Cruz, CA, USA (2012)
36. Skantze, G., Al Moubayed, S., Gustafson, J., Beskow, J., Granström, B.: "Furhat at Robotville: A Robot Head Harvesting the Thoughts of the Public through Multi-party Dialogue". In Proceedings of IVA-RCVA. Santa Cruz, CA (2012)

37. Harel, D.: Statecharts: A visual formalism for complex systems. Science of computer programming, 8(3), 231-274 (1987)
38. Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: Conference on Human Factors in Computing Systems, pp. 858-859 (2002)
39. Ba S.O., Odobez J-M.: Recognizing visual focus of attention from head pose in natural meetings. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39(1), 16-33 (2009)
40. Johansson M., Skantze, G., & Gustafson, J.: Head Pose Patterns in Multiparty Human-Robot Team-Building Interactions. In: Herrmann, G, Pearson, M.J., Lenz, A., Bremner, P., Spiers, A., Leonards, U. (eds.) International Conference on Social Robotics, Bristol, UK, LNCS, vol. 8239, pp. 351-360, Springer, Heidelberg (2013)
41. Al Moubayed, S., Beskow, J., Granström, B.: Auditory-Visual Prominence: From Intelligibilitty to Behavior. Journal on Multimodal User Interfaces, 3(4), 299-311 (2010)
42. Al Moubayed, S., Beskow, J.: Effects of Visual Prominence Cues on Speech Intelligibility. In:Auditory-Visual Speech Processing, AVSP'09, Norwich, England (2009).
43. Streefkerk, B., Pols, L. C. W., ten Bosch, L.: Acoustical features as predictors for prominence in read aloud Dutch sentences used in anns. In: Eurospeech. Budapest, Hungary (1999)
44. Bevacqua, E., Pammi, S., Hyniewska, S. J., Schröder, M., Pelachaud, C.: Multimodal backchannels for embodied conversational agents. In: The international conference on intelligent virtual agents. Philadelphia, PA, USA (2010)
45. Zhang, J. Y., Toth, A. R., Collins-Thompson, K., Black, A. W.: Prominence prediction for super-sentential prosodic modeling based on a new database. In: ISCA speech synthesis workshop, Pittsburgh, PA, USA (2004)
46. Al Moubayed, S., Chetouani, M., Baklouti, M., Dutoit, T., Mahdhaoui, A., Martin, J-C., Ondas, S., Pelachaud, C., Urbain, J., Yilmaz, M.: Generating Robot/Agent Backchannels During a Storytelling Experiment. In Proceedings of (ICRA'09) IEEE International Conference on Robotics and Automation. Kobe, Japan (2009)
47. Terken, J.: Fundamental frequency and perceived prominence of accented syllables. The Journal of the Acoustical Society of America 89, 1768-1776 (1991)
48. Wang, D., Narayanan, S.: An acoustic measure for word prominence in spontaneous speech. IEEE Transactions on Audio, Speech, and Language Processing 15, 690-701 (2007)
49. Kullback, S.: Information Theory and Statistics. John Wiley and Sons (1959)
50. Hotelling, H., Eisenhart, M., Hastay, W., Wallis, W. A.: Multivariate quality control. McGraw-Hill (1947)
51. Cheveigne, A. D., Kawahara, H.: Yin, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America 111, 1917-1930 (2002)
52. Greenberg, S., Carvey, H., Hitchcock, L., Chang, S.: Temporal properties of spontaneous speech - Asyllable-centric perspective. Journal of Phonetics 31 ,465-485 (2003)
53. Yilmaz, A., Javed, O., and Shah, M.: Object tracking: A survey. ACM Comput. Surv. 38, 4, Article 13 (2006)
54. Rienks, R., Heylen, D.: Dominance Detection in Meetings Using Easily Obtainable Features. LNCS, vol. 3869, pp. 76–86, Springer (2006)

Appendix A

Table 3. Tutor Assessment Questionnaire

Tutor Assessment Questionnaire

For the assessment of the tutor, please fill out the following questionnaire. The questionnaire consists of pairs of contrasting characteristics that may apply to the tutor. The numbers between the characteristics represent gradations between the opposites. You can express your agreement with the characteristic by ticking the number that most closely reflects your impression. Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute does not apply completely to the particular tutor. Nevertheless, please tick a number in every line.

Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression.

It is your personal opinion that counts. Please remember: there is no wrong or right answer!

Name/Nickname (for identification): _____

I think that the tutor is...

		1	2	3	4	5	6	7	
1	annoying								enjoyable
2	unfriendly								friendly
3	bad								good
4	unpleasant								pleasant
5	does not meet expectations								meets expectations
6	socially obstructive								socially supportive
7	unpredictable								predictable
8	slow								fast
9	inefficient								efficient
10	cluttered								organized
11	confusing								clear
12	complicated								easy
13	not understandable								understandable
14	acts as if he/she does not know what to do								acts as if he/she does know what to do
15	doesn't care to hold our attention								holds our attention
16	doesn't give us feedback on the quality of our work								gives us feedback on the quality of our work
17	doesn't help us with our task								helps us with our task
18	doesn't realise when we fail to agree								realises when we fail to agree
19	hesitant								never hesitates
20	if I have something to say, the tutor won't listen								if I have something to say, the tutor will listen
21	impatient								patient
22	not consistent in his/her behavior								consistent in his/her behavior
23	not motivating								motivating
24	passive								active

Appendix B

Table 4. The annotation scheme employed for the manual analysis of the tutor conversational behavior.

Annotation layers	Values
Speech_activity	Free text
Dialogue acts	Take, Accept, Grab, Offer, Complete, Yield, Hold, Back-channel
Turn management	Take, Accept, Grab, Offer, Complete, Yield, Hold, Back-channel
Feedback	Perception/Understanding (Give-Elicit) Accept (Give-Elicit), Non-accept (Give-Elicit)
Verbal_feedback	Free text
Face_general	Smile, Laugh, Scowl
Functions_Face	Feedback, Turn Management
Head_movement	Nod(s), Shake, Jerk, Tilt, Turn, Forward, Backward
Functions_ Head_movement	Feedback, Turn Management
Gaze	Attention_Person_Right, Attention_Person_Left, Attention_Object, Glance
Functions_Gaze	Feedback, Turn Management
Eyes	Wide_open, Semi-closed, Wink, Blink
Functions_Eyes	Feedback, Turn Management
Eyebrows	Raise, Frown
Functions_Eyebrows	Feedback, Turn Management
Mouth	Open, Closed
Functions_ Mouth	Feedback, Turn Management
Cards	Card id