Intel Serv Robotics (2008) 1:3–26
DOI 10.1007/s11370-006-0001-9

ORIGINAL RESEARCH PAPER

Low-level grounding in a multimodal mobile service robot conversational system using graphical models

Plamen Prodanov · Andrzej Drygajlo · Jonas Richiardi · Anil Alexander

Received: 3 September 2005 / Accepted: 20 September 2006 / Published online: 30 January 2007

Abstract The main task of a service robot with a voice-enabled communication interface is to engage a user in dialogue providing an access to the services it is designed for. In managing such interaction, inferring the user goal (intention) from the request for a service at each dialogue turn is the key issue. In service robot deployment conditions speech recognition limitations with noisy speech input and inexperienced users may jeopardize user goal identification. In this paper, we introduce a grounding state-based model motivated by reducing the risk of communication failure due to incorrect user goal identification. The model exploits the multiple modalities available in the service robot system to provide evidence for reaching grounding states. In order to handle the speech input as sufficiently grounded (correctly understood) by the robot, four proposed states have to be reached. Bayesian networks combining speech and non-speech modalities during user goal identification are used to estimate probability that each grounding state has been reached. These probabilities serve as a base for detecting whether the user is attending to the conversation, as well as for deciding on an alternative input modality (e.g., buttons) when the speech modality is unreliable. The Bayesian networks used in the grounding model are specially designed for modularity and computationally efficient inference. The potential of the proposed model is demonstrated comparing a conversational system for the mobile service robot RoboX employing only speech recognition for user goal identification, and a system equipped with multimodal grounding. The evaluation experiments use component and system level metrics for technical (objective) and user-based (subjective) evaluation with multimodal data collected during the conversations of the robot RoboX with users.

Keywords Service robots · Spoken interaction · Grounding · Bayesian networks · Efficient inference

P. Prodanov · A. Drygajlo · J. Richiardi
Perceptual Artificial Intelligence Laboratory, Signal Processing Institute, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland

A. Alexander
Clarifying Technologies Ltd, Oxford, UK

P. Prodanov (✉)
TBS Holding AG, Rösslimatte 8,
CH-8808 Pfäffikon SZ, Switzerland
e-mail: plamen.prodanov@tbsinc.com

1 Introduction

Mobile service robots are physical agents that are designed to act in the real world, using their mobility to perform tasks useful for humans. Service robots perform some fixed number of services specific to the particular service robot application. These services can be, e.g., presenting exhibits in the case of mobile tour-guide robot [4,8] or object delivery in the case of robot assistants [17]. Service robots need to interact with their users to decide on which service to perform.

Users of service robots are most of the time unprepared ordinary people, i.e., people without any prior experience with robotics. When designing a communication interface for ordinary people, intuitiveness and usability become very important. Speech is an intuitive communication means for humans and for that reason service robots often employ automatic speech synthesis and recognition in performing their communication tasks. Speech recognition technology has gained

performance in recent years that enables real-world applications [16]. Most of the service robot applications, however, take place in open spaces, where speaking people other than the user and the robot equipment itself can contribute to high levels of noise in the acoustic space. Today's state-of-the-art speech recognition techniques yield to recognition errors in noisy environments [23]. Speech recognition errors can lead to subsequent robot behavior that may not meet the user expectations and interest. In the case of such robot's behaviors, the user can stop interacting and move away at any time. The interaction hence results in a communication failure.

This paper exploits a model for reducing the risk of communication failures in a conversational system of a mobile service robot, inspired by the concept of grounding. Service robot conversation is performed in the form of spoken dialogue, where the dialogue structure can be presented as pairs of consecutive robot and user turns. After each user turn the robot has to infer the user goal (i.e., the user intention of requesting a particular service) using speech recognition in order to decide what service to perform. When designing conversational systems for service robots we have to be aware that misunderstandings about the communication goals of the participants occur even in conversations between humans that are thought to have "perfect" speech recognition abilities. If not handled, these misunderstandings might result in communication failures. In the case of a conversation between people, misunderstandings are collaboratively resolved by the dialogue participants. People coordinate their individual knowledge states by systematically seeking and providing evidence about what they say and understand, which is known as the process of grounding in conversation [5,6].

In human–computer spoken dialogue the process of grounding has been modeled using state-based models, where the states represent increasing levels of grounding between the dialogue participants [5,6]. Low levels of grounding are related to the minimal initial conditions for establishing communication. For example, spoken communication is impossible without a user who attends to the conversation. Higher levels of grounding are related to joint activities of establishing common beliefs and intentions between dialogue participants, and it requires already established low-level grounding.

Human–robot interaction presents conceptually new type of spoken interaction, in which both the robot and the user have the freedom to move. In such conditions the existing grounding models in human–computer interaction has to be fitted to the specific requirements of low-level grounding in the service-robot dialogue. Hence the focus of this paper is on providing models of low-level grounding to aid spoken communication with service robots.

The amount of effort that people spend to ground their conversation is governed by a "grounding" criterion [5,6]. The grounding criterion evaluates if the level of understanding in dialogue is sufficient for the current dialogue purpose, or if there is a risk for misunderstanding. In the state-based grounding model, the grounding criterion depends on the strength of evidence provided by the model for reaching the different grounding states. Reaching different states would signify different levels of grounding and will require corresponding grounding actions. For example, in very noisy acoustic conditions a speaker will specially seek for the attention of the listener by looking at him in the eyes, using much louder voice and repeating the important terms waiting for an appropriate acknowledgement. On the contrary, in quiet conditions all these actions might slow down the interaction and even frustrate the listener. In a similar way, a service robot managing spoken dialogue with people will need to establish sufficient level of low-level grounding with its user for minimizing the risk for communication failures. A sufficient level of grounding would mean that the robot has obtained sufficient evidence that the following grounding states have been reached: (1) user is attending to the conversation and (2) the speech modality is reliable in the current acoustic conditions.

In human–robot interaction, evidence for reaching grounding states can be delivered by information from speech modality as well as other input modalities available on the robotic platform. For example, the state that the user is attending to the conversation can be revealed by the detection of a frontal face in the video modality, as well as a particular legs-pattern in the laser scanner modality. In the case of adverse acoustic conditions detected by the speech input modality, the robot can ask the user for a repeated trial in which alternative input such as buttons can be used. In that way the robot can avoid the unreliable speech recognition in very noisy conditions. To ensure such functionality the robot needs a model to infer the corresponding grounding states such as the state of attending user or the state of speech modality reliability. Since the end-users' behavior can vary largely during their communication with the robot and the acoustic conditions are a priori unpredictable, the corresponding grounding states can be never inferred with certainty. Moreover, the limitations of the current sensor technology that is prone to measurement errors can lead to imprecise modality information. Hence, models based on deterministic mapping between input modality features and corresponding grounding states and user goals can lack

sufficient robustness to the uncertainties of real-life service robot dialogue. Probabilistic models can deal with uncertainty using parametric models of distributions over random variables. The random variables can be associated with the grounding states and features derived from the robot modalities. The relations between the grounding states and their corresponding modality features can be seen as causal relations. Bayesian networks are widely accepted framework for efficient modeling of the probability distribution over a set of random variables by encoding the independence assumption behind the variables' causal relations. Hence, we use Bayesian networks for grounding modeling of spoken interaction between a user and a mobile service robot in mass exhibition conditions (tour-guide robot).

While incorporating information from additional modalities can bring benefits [33] in detecting possible communication failures during interaction, the resulting model that should infer grounding states and user goals using Bayesian networks can become complex and computationally expensive. Hence, providing Bayesian network topologies that allow straightforward incorporation of new modalities in the grounding model and computationally efficient inference becomes important.

The paper is structured as follows. In Sect. 2 the model of grounding for service robot dialogue is motivated from related work in the fields of cognitive science and human–computer interaction. In Sect. 3 we define a multimodal grounding architecture with four grounding states evaluated in two phases of grounding during human–robot interaction. Bayesian networks are then introduced (Sect. 4) and used to model the two phases of grounding (Sect. 5), providing probabilities for reaching each grounding state. These probabilities serve as measures for grounding and identification of the user goal for the purpose of service-oriented dialogue. In Sect. 6 the proposed state-model is evaluated in experiments with data gathered during the real interactions of the service robot RoboX who serves as a tour guide robot in the Autonomous Systems Laboratory at EPFL. Finally, the potential benefits of the multimodal grounding architecture for error handling in spoken dialogue with service robots are outlined in the Discussion and Conclusion parts of the paper (Sects. 7 and 8).

2 Related work

2.1 Grounding for error handling in human–computer interaction

In the grounding theory the model of dialogue error handling is represented as an incremental process of establishing a common ground between the participants in the conversation (e.g., the user and the robot in our case). The common ground is related to the state of achieving sufficient understanding between the participants for the purpose of the conversation. In a collaborative dialogue setting, the state of sufficient understanding is closely related to the evidence that what is being said by the speaker is understood by the listener(s) under the current purpose of the conversation. Such evidence is provided by explicit and implicit feedback between the participants in the conversation. The feedback can be negative— signaling misunderstanding and contributing to a decreased level of understanding, or positive—signaling increased level of understanding and finally agreement. Based on the lack or presence of sufficient evidence of mutual understanding, people employ grounding actions. In human–computer spoken interaction the recognition error correction dialogues can be seen as such actions [3].

Table 1 Unimodal state model of grounding in conversation

State	Description
State 0	*R* did not notice that *U* uttered any *u*
State 1	*R* noticed that *U* uttered *u*
State 2	*R* correctly heard *u*
State 3	*R* understood *u*

In their seminal work Clark and Schaefer [6] introduced a state model to represent the incremental process of grounding in a collaborative conversation between dialogue participants. In this model the level of sufficient understanding is explicitly represented by a set of states that an addressee *R* attributes to a speaker *U* and an utterance *u*. The state model is depicted in Table 1.

All the states have to be reached in order to consider the current participant dialogue contribution as grounded. Whether a state has been reached depends on the evidence provided by the feedback from the speaker as well as by environmental factors related to acoustic noise. The need for grounding actions arises whenever *R* has failed to reach one of the states in the model. In the case of a human–computer dialogue the speech modality should provide all the evidence for inferring the four grounding states in Table 1. Therefore, we refer to this model as the unimodal grounding model. The unimodal grounding model was further extended by Traum et al. [38,39] who have proposed the Conversational/Grounding Acts model, contributing to the taxonomy of speech acts with grounding-related acts. The authors proposed quantitative model for the utility of a grounding acts, based on the value of a grounding criterion measure, the added effect of the grounding act and its cost. Other authors [3] have also extended

the original grounding model with additional states and related grounding actions, commenting on the effect of the grounding criterion on selected grounding actions.

All the above studies concentrate on grounding using only speech as a communication medium. Their grounding models and definitions for the grounding criterion measure provide only specific solutions to the particular study.

2.2 Bayesian networks for grounding in spoken interaction

Horvitz and Paek [15,30] have proposed a computational model for the process of unimodal grounding motivated initially by the Clark and Shaefer architecture. In this model they regard grounding and error handling in dialogue as a process of making decisions under uncertainty in a four-level architecture called the "Quartet". The uncertainty in taking a decision can arise from the unreliable speech recognition results under noisy conditions, the inherent ambiguity in the way humans express themselves in conversation, etc. The uncertainties in the four-level grounding state inference (channel, signal, intention and conversation) are modeled using Bayesian networks. The cost of grounding (grounding criterion) and subsequent cost of the grounding actions is modeled using decision networks (influence diagrams) that are essentially extended version of Bayesian networks. The authors have applied the method in three different dialogue systems — the Bayesian Receptionist [13], the Presenter [31] and the DeepListener [14]). The Receptionist is handling typical services offered by receptionists at Microsoft campus. For this purpose the system is able to detect a fixed number of user goals and map them to desired services. The presenter is a voice-driven presentation system that is able to detect only voice commands related to the slide manipulation. The DeepListener is a command and control system.

The model of Horvitz and Paek is influential in that it provides computational model for unimodal grounding and error handling in dialogue based on identifying user goals and providing appropriate services. The authors give details on how such a system can be built by providing the Bayesian networks involved in the "Quartet" model. However, the intuition behind building the necessary topologies is not stated explicitly. The networks used seem monolithic, densely connected with multiple layers. Such type of Bayesian networks are difficult to interpret and reuse in other systems, since authors do not provide guidelines on how they were composed. Densely connected and multi-layered Bayesian networks are also known to be computationally expensive as far as probabilistic inference is concerned [7,22].

All of the grounding-based error handling models presented till now are oriented toward extracting information mainly from one input modality, i.e., the speech modality. In human–robot interaction, however, the speech modality can fail to provide sufficient information in order to avoid typical communication failures, such as the one resulting from a user that has abandoned conversation. In noisy acoustic conditions the speech recognition can still process background noise and infer a valid user goal leading to "awkward" behavior from the side of the robot. In such conditions available modalities utilized by the robot for other purposes (e.g., navigation) such as laser and video provide additional information to be used in the grounding model. For example, the lack of a user as detected in the laser scanner reading can point out recognition errors that could otherwise result in valid user goals. The above observations outline the need for adapting and extending the initial states of the grounding model in Table 1 with new states associated with the different robot modalities (Sect. 3).

In the following section we investigate error-handling techniques in conversational systems of service robots, focusing especially on the use of the concept of grounding.

2.3 Grounding in human–robot interaction

The need of a systematic way of seeking and providing user feedback, during human–robot interaction, is one of the main motivations behind the use of grounding models [17]. In [17] grounding, i.e., establishing common knowledge of a dialogue topic is seen as very important prerequisite for sustaining successful human–robot communication. In this study, grounding is defined at low and high levels of interacting. For high level grounding, the speech modality on the robot is used to extract information about the intention of the user. The robot in the study is a service robot, assisting a handicapped person in her/his everyday needs. In particular, the robot was designed to deliver objects to different locations (e.g., cups in the kitchen). The high level grounding is responsible for resolving ambiguities in user goal identification, when using natural spoken input to specify the robot tasks. The user goals can be related to one of the two possible tasks (*Go to mission* and *Deliver mission*). Each of these tasks needs predefined pieces of information (e.g., location in the *Go to mission* location and object specification in the *Deliver mission*). Since some of the information could be missing or skipped in the spoken user input, grounding actions are used

such as clarification questions to resolve the resulting ambiguity. The low level grounding, on the other hand, is dedicated to providing gestural feedback to the user through a small physical human-like character (CERO).

In [2] the robot is equipped with a layered attentional system that is responding to high-level events related to interaction (e.g., missing concepts in conversation) as well as low-level events (e.g., high level of acoustic noise). The authors argue that combining low and high level feedback to the user about the state of the robot results in more intuitive human–robot interaction. A process similar to grounding is also discussed in [36]. The authors describe engagement rules in interaction with a static penguin-like robot Mel. They describe techniques very similar to the process of incremental grounding without explicitly referring existing work such as [6]. Instead, they motivate their interactive engagement system from their user studies.

Both the studies [2] and [17] describe grounding from the perspective of high-level feedback-provision during dialogue (goal clarification level). The grounding is performed using only the speech modality. However, the setting of human–robot interaction with mobile service robots differs from the more general case of human–computer interaction in that the user is free to move like the robot. User may also leave the robot at any time. Therefore, it is important that before providing high-level grounding actions the robot detects the state of the user attendance in the process of interaction.

2.3.1 Exploiting different input/output robot modalities

Detecting user activity is the purpose of the robot attentional system [25]. This system can be seen as the component providing the robot with user and situation awareness. Situation awareness is the process that identifies entities in the surrounding environment that are essential for the process of human–robot interaction.

The robot attentional systems often employ multimodal solutions and can provide information for low-level grounding. In [29] the authors utilize audio–visual approach for people tracking in the attentional system of the robot SIG—a stationary humanoid upper torso. In a later study [37], depending on people's distance and activity, SIG is also able to classify users to different "friendliness" states, incorporating information from speech, video and tactile input modalities. In [24] laser and video are used in the attentional system for detecting and tracking people in human–robot interaction with the mobile service robot Biron. In a follow-up study about the same robot [26] the authors describe a multimodal (human-style) interaction system for the robot Biron, who has to learn new objects in the home of its user. The robot uses a multimodal interface based on speech and deictic pointing gestures for object specification. Authors introduce grounding on the higher interaction level of disambiguating the spoken input through clarification questions.

All the above studies concerning grounding in human–robot interaction are focused on high-level grounding in dialogue, relying on information derived from the speech modality. However, low-level grounding feedback from the side of the user, such as the state of attendance to the conversation, can reveal very common situations that can produce recognition errors. Detecting the state of user attendance to the conversation would require additional information from modalities complementary to the speech modality. The attentional system of the robot can provide such information to the process of grounding in human–robot interaction.

Finally, to enable low-level multimodal grounding, techniques for multimodal signal fusion need to be used. In the sections that follow we present a Bayesian network based model for combining multiple modalities in an extended grounding model for speech-based interaction with a service robot. Our main objective is to define special structure in the Bayesian network models that will allow modularity in the process of adding new modalities as well as computationally efficient inference.

3 Multimodal grounding in service human–robot interaction

To build the grounding model for speech-based interaction with a service robot we take inspiration from the state model after [6] presented already in Table 1.

3.1 Grounding states in human–robot interaction

We adapt the original model with the states needed by a "collaborative" service robot in order to decide that the input audio signal is sufficiently grounded relying on information from speech and non-speech modalities. The updated multimodal grounding state model is depicted in Table 2. To avoid interpreting background noise as user input the service robot has to be able to detect the potential user from people that are not using the system. It should have positive feedback from the user for reaching states S0 and S1 in Table 2. Interested and collaborative users provide positive feedback showing attention through looking at the robot. To facilitate collaborative communication, the devices of the service robot are typically arranged to mimic anthropomorphic elements (e.g., a mechanical face), where a camera is typically located (Fig. 1). A collaborative user is assumed to

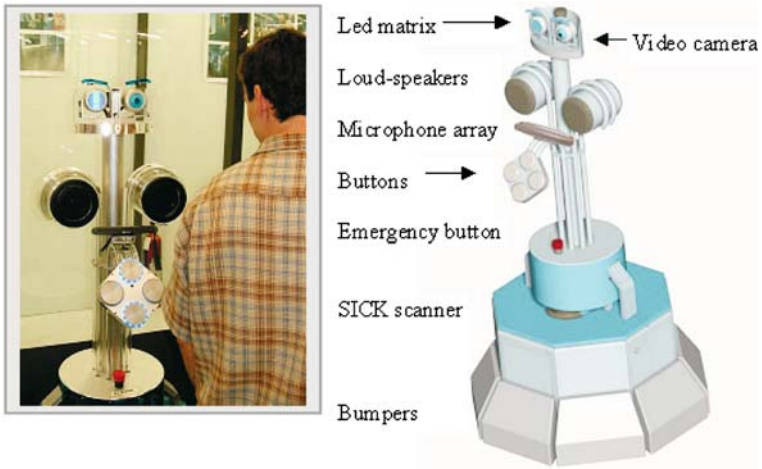


Fig. 1 The mobile service robot RoboX

Table 2 Multimodal state model of grounding in human-robot conversation

State	Modality/Event	Description
S 0	Laser/UR = 1	***U***ser present in ***R***ange for communication
S 1	Video/UA = 1	***U***ser ***A***ttending (looking at the robot)
S 2	Speech/SMR = 1	*S*peech ***M***odality is ***R***eliable
S 3	Speech/UG ≠ 0	Robot identified a valid ***U***ser ***G***oal

*UR U*ser in *R*ange, *UA U*ser *A*ttending, *SMR* Speech *M*odality *R*eliability, *UG U*ser *G*oal

stay close to the robot (S0 reached) looking at the robot's "face" (S1 reached) while communicating her/his user goal. A correct user goal interpretation using speech recognition requires that the speech recognition result is reliable (S2 reached), where speech recognition reliability is mostly affected by the level of the background acoustic noise [16]. To be understood by the robot the user request has to be interpreted as a valid user goal, i.e., a goal that can be mapped into an existing service offered by the robot (S3 reached). Similar to the original model, reaching all the states in Table 2 will signify that the user speech input is grounded (understood by the robot) for the purpose of the service robot task oriented dialogue.

Failure or success to reach a given state is signaled by evidence provided in the information from the robot's input modalities. The robot input modality can be defined as one of the main input data channels provided by the different robot sensors, such as speech, video, laser, etc. Information is extracted out of each modality in the form of events that can be inferred from the raw modality data. For example, the binary event "UR = 1" that a user is staying in close range in front of the robot can be inferred from the information contained in the laser scanner data. The binary event "UA = 1" — "*U*ser attending" can be inferred from information extracted from the video modality for a presence of a frontal face in the camera view. The event "SMR = 1" corresponding to "speech modality is reliable" can be inferred from information from the speech modality and the level of acoustic noise in particular. SMR = 0 means that there is error at the output of the speech recognition (see Sect. 5.2 for more details). Finally, the speech modality is used to interpret the user goal defined by the event UG, where UG = 0 means an undefined and UG ≠ 0 means a "valid" user goal, i.e., goal that can be mapped onto existing robot-provided service. Examples of valid user goals are presented in Sect. 3.2.1. The above events and their association with the grounding model states are depicted in Table 2.

Whether a grounding state is reached will directly depend on the strength of evidence for the above events as provided by the information from the input modalities data. Given that the last grounding state is reached (UG ≠ 0) would mean that S2 has been reached too (SDR = 1), which in turn would mean that S1 is reached (UA = 1) and S0 is reached (UR = 1), since an attending user implies a user who is close to the robot. All the above states and the propagation of evidence about their possible instantiations can be modeled by a Bayesian network. Then the strength of evidence about the modality related events can be quantitatively estimated by the posterior probability of the event given the evidence from the modalities data, for example the posterior probability for an "undefined user goal": $P(\mathrm{UG} = 0|E = e)$, for the variable UG in the Bayesian network given the evidence $E = e$ from the input modalities. The posterior probabilities over the grounding states can be used as grounding criterion in the case of service robots. Their values signify possible failures to reach a

particular state in the grounding model that will require corresponding grounding actions.

In building the grounding model for service robot dialogue we use the mobile tour-guide service robot RoboX (Fig. 1) as an example. RoboX was designed to provide tour-guiding services and was successfully deployed at the Swiss National Exhibition (Expo.02) in 2002 [20]. For the purpose of providing interactive tours, RoboX (Fig. 1) is equipped with the following modalities: speech input modality (recognition), interactive buttons and video camera as input modalities, and LED matrix animations, expressive face (moving eyes and eyebrows), speech output modality (synthesis) as output modalities. For the tasks of navigation and obstacle avoidance the robot is additionally equipped with the following input modalities: two laser scanners (laser range finders SICK), emergency stop button and bumpers for avoiding collision with obstacles that cannot be detected by the laser scanner beam [18,19].

3.2 Multimodal interaction with service robots

Service robot dialogues can be generally defined as a sequence of turns. Each pair of dialogue turns contains verbal interaction in the form of initiative/response (e.g., robot's question/visitor's answer) pair, during which the speech recognition is typically used to infer the "goal" of the speaker in the context of the current turn. The response part of the initiative/response pair is initiated by a phase of multimodal input acquisition and is concluded by a phase of a robot multimodal response. During the phase of the input acquisition each input modality operates on features extracted from the input modality data to infer corresponding events related to the input modality features. The events and the feature values are typically represented by discrete and continuous variables. The possible event values have well defined meaning in the service task oriented dialogue (Table 2). The response phase in the dialogue turn employs one or more of the robot output modalities in performing the robot service. The output modality can be defined as one of the main output functionalities of the robot through which the external environment is manipulated. The combined output from these modalities can be seen as the actions that the robot performs in fulfilling its services.

3.2.1 Identifying user goals

The speech input modality of RoboX can recognize spoken keywords. The keywords currently used can be directly mapped into user goals corresponding to services offered by the robot. The full sequence of service dialogue turns in the case of RoboX is typically defined in advance according to the exhibition cite plan. We assume that the spoken utterances (keywords) coming from users during interaction can be mapped into a finite number of turn dependent user goals, which are used to infer the next dialogue turn. Then the key issue in spoken service dialogue management is to decide on the most likely user goal into the current dialogue turn.

At Expo.02 one complete tour consisted of five exhibit presentations [19]. During the exhibition RoboX interacted with individual visitors as well as crowds of people in very noisy acoustic conditions. In these conditions the robot's ability to attract and keep people involved in the interaction was very important for his success as a tour-guide. On the other hand, the tasks that most tour-guide robots are expected to perform typically requires only a limited amount of information from the visitors. Therefore, we have chosen a very limited but meaningful speech recognition vocabulary. The solution adopted was based on yes/no questions initiated by the robot and yes/no answers as a meaningful universal commands. The dialogue turn in the case of RoboX was at the beginning of each exhibit's presentation and consisted of yes/no question from the robot and answer from visitor (e.g., the tour-guide robot asks the visitors if they want to see the next exhibit). Successful speech recognition can be then measured by the average number of correctly recognized responses at the beginning of each exhibit presentation.

The initial experiments during Expo.02 showed that such an interaction scheme could be seriously challenged by the visitors' behavior. There were often cases when people did not follow the choice suggested by the robot [8], using out of vocabulary words and even giving both yes and no answers or simply remaining silent. Therefore the speech recognition system of RoboX was designed to distinguish between the keywords yes, no and out-of-vocabulary words, fillers, coughs, laughs and general acoustic phenomena different from the keywords called garbage words (GB). The Observed Recognition Result $\mathrm{ORR} = \{\mathrm{yes}, \mathrm{no}, \mathrm{GB}\}$ is then mapped into three possible user goals (UG), accounting for the visitor intention: "the user is willing to see the next exhibit" ($\mathrm{ORR} = \mathrm{yes}$ then $\mathrm{UG} = 1$); "the user is unwilling to see the next exhibit" ($\mathrm{ORR} = \mathrm{no}$ then $\mathrm{UG} = 2$) and "user goal is undefined" ($\mathrm{ORR} = \mathrm{GB}$ then $\mathrm{UG} = 0$).

In its present state the speech modality of RoboX has been extended to cover more keywords and user goals than just the yes/no pair. In general, the answer of the user can contain a keyword used as a command to request one of $N - 1$ possible services or can be undefined, corresponding to the garbage word (GB) (Fig. 2). We have currently limited the number of possible goals

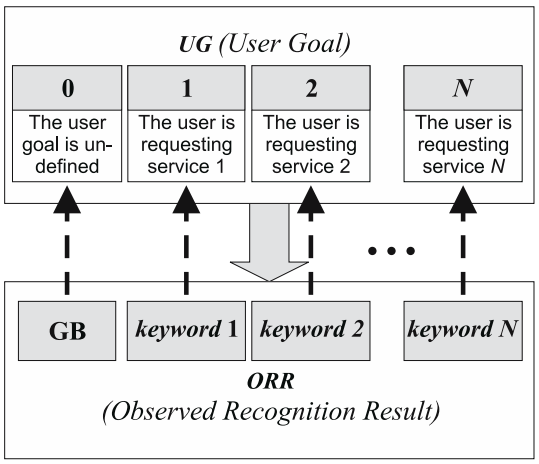


Fig. 2 ORR to UG mapping

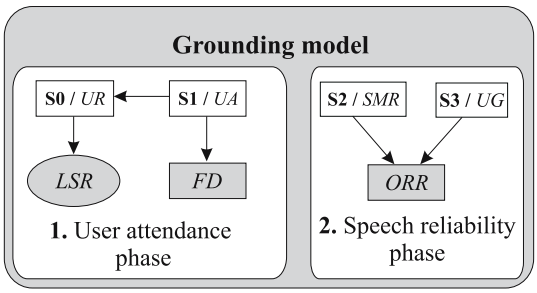


Fig. 3 Two-phase grounding architecture for reliable speech-based UG identification

per turn to $N = 3$ in order to improve the robustness of speech recognition to substitution errors. In this case we define three possible user goals UG = 1, first possible service UG = 2, second service and UG = 0, undefined user goal at each dialogue turn. The concrete user goal and service definition depends on the particular turn in the dialogue. The different meanings for the UG used in our experiments are described in detail in Sect. 6.

3.3 Two-phase grounding for user goal identification

The speech modality is the main modality used for inferring the goal of the user out of the possible goals defined at each particular dialogue turn. The User Goal (UG) is derived from the spoken user request for a service during the input acquisition phase. To account for the cases when the user goal cannot be interpreted into the set of defined goals we include an undefined user goal at each dialogue turn. The undefined user goal often results from communication failures, such as in the case when out of the robot vocabulary words are used by the user in answering to the robot or when the user has left in a middle of a conversation answering to other people calling her/him. In order to minimize the possible communication failures, user goal inference is performed in two consecutive phases in the multimodal grounding model.

- In the first phase the robot requires sufficient level of grounding as far as the user attendance to the conversation is concerned. Sufficient level of grounding requires strong evidence that the state S1 is reached, which also implies that S0 is reached (Table 2). This is needed for the robot to proceed to the second phase.
- In the second phase, the robot seeks for sufficient level of grounding as far as the speech modality reliability is concerned. This would mean that state S2 is reached, after which S3 can be evaluated from the speech recognition result.

The reason behind the phase definition stems from the fact that it does not make sense to check the modality reliability and infer a user goal, if the user is not there, or is not paying the needed attention in the conversation. In that cases the user goal UG can be set to the undefined goal (UG = 0). Only after achieving the two phases of grounding, the robot can reliably identify user goals from the underlying speech modality. The outlined phases for inferring user goals are depicted in Fig. 3.

The grounding states and their associated modality events are depicted in the figure along with arcs indicating the causal relations between them as well as the corresponding modality features. LSR denotes the laser scanner reading, which is the raw feature supplied by the laser modality. FD denotes the state of face detection that is a binary feature derived from the video modality. ORR corresponds to the observed recognition result (recognized keywords) supplied by the speech modality. In Fig. 3, the modality-specific events (e.g., UR, UA, SMR, UG) can be seen as the causes behind the particular input observations (feature values — LSR, FD, ORR). Through its events every distinct modality provides information about a particular aspect of the user goal. The final user goal can be causally related to specific instances for all modality specific events. For example, a valid user goal (UG $\neq$ 0) would be the cause for UR = 1 and UA = 1. Inferring the correct user goal can be possible only when fusing information from more than one of the input modalities. Thus, fusing the different user goal aspects, as represented by the possible instantiations of the modalities' events can result in more robust user goal identification, compared with using only one modality [33]. In the fusion schema, we have to take into account the fact that the events detected by each modality are not deterministically related with the underlying modality features. For

example, the recognition result (ORR) is affected by the level of acoustic noise as well as the speech variability from user to user resulting into different versions of the same underlying UG. Hence, the cause-effect relation between the user goal and the speech recognition result should be seen as probabilistic. This argument is valid for the other modalities as well, i.e., laser and video. The uncertainty in this case can result from measurement errors as well as imperfect detection algorithms. In that case, the influence of the cause can be modeled through a conditional distribution over the set of outcomes of the resulting event. Bayesian networks have been shown to perform inference about probabilistically related events compatible with the notion of causal reasoning [21].

4 Bayesian networks

Bayesian Networks (BNs) are graphical models used to describe a joint probability distribution (pdf) over a finite set of random variables [32]. The pdf structure is characterized by a directed acyclic graph (DAG) in which the nodes represent random variables and the lack of arcs represents conditional independence assumptions between the variables. The variables can be discrete and continuous. They have well defined meaning in the particular problem domain, modeled by the network. The BN's topology is often built a priori on the basis of knowledge of intuition [21]. In all Bayesian networks in this paper we use rectangles to represent discrete and ovals to represent continuous variables, shading marks observed variables. After defining the model variables the causal relations between them should be considered. These relations are represented by the arcs' direction. The arcs point from all parent variables to their children variables.

From the probabilistic point of view the arcs converging at a given node specify the conjunction of all variables that appear as conditioning ones (parents) for the node's conditional probability distribution (CPD) term. Hence, a BN is completely defined by the triple (V, A, CPD), where V is the set of nodes associated with the random variables, A is the set of arcs and CPD is the set of conditional probability distributions associated with the nodes' variables. The CPDs can be tables in the case of discrete variables. In this paper we use single Gaussians for the continuous ones.

4.1 Inference in Bayesian networks

The basic task of probabilistic inference in Bayesian networks is to compute the posterior distribution for a set of query variables, given some observed event, i.e., an evidence for some observed (evidential) variables. Formally, we calculate $P(X_Q|E)$, where $X_Q \in X$ is the subset of query variables from the full set of unobserved variables $X = \{x_0, \ldots, x_{L-1}\}; E = \{e_0, \ldots, e_{M-1}\}$ is the subset of the observed (evidential) variables and $V_N = X \cup E = \{v_0, \ldots, v_{N-1}\}$ is the set of all N random variables in the Bayesian network. Once the conditional probability distribution functions for all the nodes given their parents are defined, an exact or approximate inference on each node in the network can be done [28,32]. In the simplest and least efficient case exact inference can be performed through marginalizing the full joint pdf after entering the particular observed value (the evidence e) for the observed variables $E = e$:

$$P(X_Q|E = e) = \alpha \cdot P(X_Q, E = e) = \alpha \cdot \sum_{X \setminus X_Q} P(V, E = e), \tag{1}$$

where $P(X_Q, E = e)$ is a set of values for all possible X_Q values and α is the normalization constant needed to make sure that the entries for $P(X_Q|E = e)$ sum up to 1. Note that, taking into account the particular observed value ($E = e$) the term $\alpha = 1/P(E = e)$ remains constant for the set of values for X_Q and can be seen as a normalization constant. In that sense it is more efficient to use the already calculated $P(X_Q, E = e)$ values and simply normalize them, so that the sum of the final entries is 1 [35]. $X \setminus X_Q$ denotes set subtraction, i.e., the summation is over all possible values for the unobserved (non-evidential) variables that are in the set X and are not in the set X_Q. Then in order to perform consistent inference, estimates for the conditional probability distribution parameters have to be learned from training examples for the network variables (the conditional probability tables for the discrete variables and the parameters of the Gaussian pdfs for the continuous ones). In the case of full observability of the variables in the training set, the estimation can be done with random initialization and a maximum likelihood (ML) training technique. During training the CPD parameters are calculated in order to maximize the likelihood of the model with respect to the training data examples (Appendix C.2 in [28]).

4.2 Decision making using inference

Finally, the inferred posterior distribution $P(X_Q|E)$ for the query variable X_Q can be used for making decisions on a particular value for X_Q, based on the observed evidence $E = e$. If X_Q is a discrete variable this last step can be seen as a classification problem in which X_Q is the

classification variable. Different optimality criteria for assigning X_Q to one of its possible class values exist. To keep the classification error at minimum we apply the maximum a-posteriori rule using an argmax criterion on the corresponding posterior probabilities:

$$\hat{x_q} = \arg\max_{x_q}(P(X_Q = x_q|E = e)). \quad (2)$$

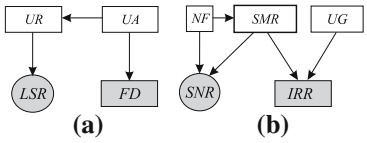
For example, in the case of the binary event UA, (X_Q = UA) applying the argmax criterion will select the UA value that results in the bigger probability out of the two possible posterior probabilities $P(\text{UA} = 0|E = e)$ and $P(\text{UA} = 1|E = e)$. This is also equivalent to establishing a threshold of 0.5 on the probability $P(\text{UA} = 1|E = e)$ when selecting between UA = 1 and UA = 0.

5 Bayesian networks for grounding

In this section we will use Bayesian networks for building the two-phase grounding model for user goal identification in service robot dialogue (Fig. 3).

5.1 Bayesian network for the attendance grounding phase

The Bayesian network for the first phase of grounding is depicted in Fig. 4a. It contains two discrete variables UR and UA corresponding to the events "User in range" for communication and "User attending" associated with the grounding states S0 and S1. These variable have direct causal impact on corresponding features derived from the laser and video modality that are represented by the two observed variables LSR and FD. LSR is a continuous variable corresponding to the laser scanner reading. Each reading contains samples within range of 360° with precision of 1°. The samples correspond to the distances from obstacles that reflects the laser beam or to the nominal range of the laser range finder which is 9 m. In order to extract features for detecting legs in the sequence of distance samples certain preprocessing steps are needed. Details concerning the preprocessing step performed on LSR for leg-detection can be found in Sect. 6. FD is a binary variable corresponding to a video modality feature indicating a face detected in the video stream. Finally, the event of "User attending" (UA = 1) to the conversation is seen as the cause of the event "User present" (UR = 1). In this case, the full set of variables is V = (UA, UR, LSR, FD), and taking into account the arcs defined in Fig. 4a, the joint pdf over V can be written as



Acronyms summary: *UR* - *U*ser in *R*ange, *LSR* -*L*aser *S*canner *R*eading, *UA* - *U*ser *A*ttending, *FD* - *F*ace *D*etected, *UG* - *U*ser *G*oal, *SMR* - *S*peech *M*odality *R*eliability, *NF* - *N*oise *F*actor, *SNR* - *S*ignal-to-*N*oise *R*ation, *IRR* - *I*nterpreted *R*ecognition *R*esult.

Fig. 4 Attendance phase (**a**) and speech reliability phase (**b**) BNs

$$P(V) = P(\text{UA})P(\text{FD}|\text{UA})P(\text{UR}|\text{UA})P(\text{LSR}|\text{UR}). \quad (3)$$

Sufficient level of grounding regarding the first grounding phase is guaranteed by UA = 1. The criterion for engaging in a grounding action at this phase is based on the posterior probability $P(\text{UA} = 1|E)$, where the set of observed (evidential) variables contains LSR and FD in this case, i.e., $E = \{\text{LSR}, \text{FD}\}$. Given the BN topology, the posterior distribution over the binary variable UA is calculated by the formula:

$$\begin{aligned}\mathbf{P}(\text{UA}|\text{lsr}, \text{fd}) &= \alpha \sum_{\text{UR}} P(\text{UA})P(\text{fd}|\text{UA})P(\text{UR}|\text{UA})P(\text{lsr}|\text{UR}) \\ &= \alpha P(\text{UA})P(\text{fd}|\text{UA}) \sum_{\text{UR}} P(\text{UR}|\text{UA})P(\text{lsr}|\text{UR}),\end{aligned} \quad (4)$$

where $\mathbf{P}(\text{UA}|\text{lsr}, \text{fd})$ denotes a two component vector, and $e = \{\text{lsr}, \text{fd}\}$ corresponds to the particular instantiations for the evidence variables LSR and FD. Particular UA value is chosen applying the argmax criterion (Eq. 2) on the posterior probabilities defined by Eq. 4.

5.2 Bayesian network for the speech reliability grounding phase

In the second phase of grounding the final user goal is inferred after ensuring sufficient speech modality reliability. The level of sufficient speech modality reliability is governed by the probability of the event of mismatch between the true user goal UG value and the one inferred from the observed recognition result (ORR). We will denote the user goal value inferred from the ORR as IRR (interpreted recognition result). Given the definitions provided in Sect. 3.2.1, we can write that if ORR = GB then IRR = 0 if ORR = keyword1 then IRR = 1, if ORR = keyword2 then IRR = 2, etc. For example in the case of ORR = {GB, yes, no}, IRR = {0, 1, 2}. Then, the event of mismatch between UG and IRR can be written as (UG ≠ IRR). To define the reliability measure we introduce a binary variable SMR,

where SMR = 1 represents the event "speech modality is reliable" (UG = IRR) and SMR = 0 represents the opposite event, i.e., (UG ≠ IRR). The Bayesian network in Fig. 4b depicts a causal model for the variables UG, IRR and SMR. In this network the user goal value can be seen as the cause of the particular interpreted recognition result, and the speech modality reliability can be seen as an alternative cause that might also point at errors in the IRR value. For example, IRR = 1 can be explained by UG = 1 and SMR = 1 (the modality makes a correct decision because the modality is reliable) or UG ≠ 1 and SMR = 0 (the speech modality is unreliable). Since the variables UG and SMR are not observable during the conversation with the robot, we need to provide additional sources of information that can be observed and can provide evidence in favor of particular (UG, SMR) values. The noise factor (NF) corresponding to the event of high level of acoustic noise can have strong causal impact on the SMR variable. A signal quality measure can be used to provide evidence for the NF variable. For example, the signal-to-noise (SNR) ratio of the speech signal can be used to account for the level of acoustic noise in the speech modality, which is known to be one of the main degradation factors for the performance of the speech recognition systems. Therefore, we define the the variables NF = {1, 0} corresponding to the binary event of "high/low level of acoustic noise" and the continuous variable SNR. SMR, can be also seen as a cause for particular SNR values. Given the BN variables set $V = (\text{UG}, \text{SMR}, \text{NF}, \text{IRR}, \text{SNR})$, and taking into account the arcs defined in Fig. 4b, the joint pdf over V can be written as

$$P(V) = P(\text{UG})P(\text{NF})P(\text{IRR}|\text{UG},\text{SMR})P(\text{SMR}|\text{NF}) \times P(\text{SNR}|\text{SMR},\text{NF}). \quad (5)$$

The posterior $P(\text{SMR}|\text{IRR},\text{SNR})$ is the distribution of the modality reliability measure. Following the network topology the posterior distribution over SMR can be written as

$$\begin{aligned}&\mathbf{P}(\text{SMR}|\text{irr},\text{snr})\\ &= \alpha \sum_{\text{UG,NF}} \Big(P(\text{UG})P(\text{NF})P(\text{irr}|\text{UG},\text{SMR}) \times P(\text{SMR}|\text{NF})P(\text{snr}|\text{NF},\text{SMR}) \Big)\\ &= \alpha \sum_{\text{UG}} \Big(P(\text{UG})P(\text{irr}|\text{UG},\text{SMR}) \times \Big(\sum_{\text{NF}} P(\text{NF})P(\text{SMR}|\text{NF})P(\text{snr}|\text{NF},\text{SMR}) \Big) \Big),\end{aligned} \quad (6)$$

where {irr, snr} correspond to the particular instantiations for the evidential variables in the Bayesian network. In the second row we apply the distributive law in order to avoid unnecessary computations [1]. We have defined the event SMR as the indicator of the event (UG = IRR). Then, given that SMR = 1 the probability values for $P(\text{IRR} = \text{irr}|\text{SMR} = 1, \text{UG})$ become $P(\text{IRR} = \text{irr}|\text{SMR} = 1, \text{UG} = \text{irr}) = 1$ and 0 for the rest UG values. In this case the Eq. 6 can be simplified in the following way:

$$\begin{aligned}&P(\text{SMR} = 1|\text{irr},\text{snr})\\ &\propto P(\text{UG} = \text{irr}) \sum_{\text{NF}} P(\text{NF})P(\text{SMR} = 1|\text{NF})P\\ &\times(\text{snr}|\text{NF},\text{SMR} = 1),\end{aligned} \quad (7)$$

where ∝ is the proportionality symbol. Since all the entries for the probabilities $P(\text{IRR}|\text{SMR} = 1, \text{UG})$ are zero except the case of UG = IRR the first summation in Eq. 6 is in fact not needed. This leads to a reduction in the number of operations needed for computing $P(\text{SMR} = 1|\text{irr},\text{snr})$. The above formula shows that the probability for reliable speech modality given values for the observed recognition result and the SNR is proportional to the prior probability of the user goal value corresponding to the particular observed IRR value multiplied by a weighted sum of two gaussian components. These components correspond to the likelihood of the observed SNR given the noise factor value and SMR = 1. The likelihood is weighted by two weight components, i.e., (1) $w_1 = P(\text{NF})$ — the prior probability of each NF value (the prior probability of high level of noise) and (2) $w_2 = P(\text{SMR} = 1|\text{NF})$ — the causal impact of the noise factor on the event (UG = IRR). The likelihood $P(\text{snr}|\text{NF}, \text{SMR} = 1)$ can be also seen as a measure of the strength of evidence for noise after observing the acoustic environment (the current SNR). To choose a SMR value we apply again the argmax criterion in Eq. 2 with $X_Q = \text{SMR}$, $E = \{\text{irr}, \text{snr}\}$.

6 Experimental evaluation

The experimental evaluation is done on two levels, i.e., component and system levels, using technical (objective) and user-based (subjective) methods. On the component level the technical evaluation is done by using accuracies as objective measures of the performance of the grounding model and the resulting performance of the user goal identification after each user turn in dialogue. The benefit of the proposed error handling framework is demonstrated by comparing the accuracy of a baseline interactive system employing only speech recognition

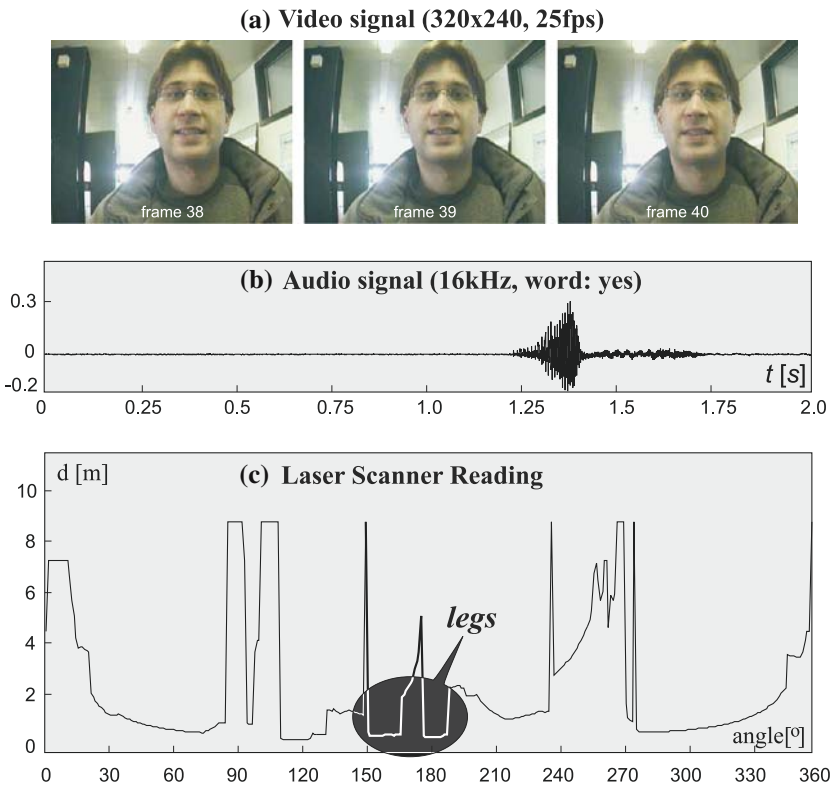


Fig. 5 Video (**a**) Audio (**b**) and Laser (**c**) modality signal

for user goal identification and a system equipped with a multimodal grounding architecture.

On the system level, the technical evaluation is done with quantitative success criteria motivated by the tour-guide robot task requirements. Finally, results from subjective usability tests are compared with the results from the technical evaluation.

The process of interactive system evaluation is initiated with a characterization step in which the particular system and components under evaluation are defined [10,11].

6.1 Interactive system characterization

To test the proposed grounding architecture model (Fig. 4) we use the mobile robot RoboX as a tour-guide in the Autonomous System Laboratory at EPFL. In the case of the tour-guide robot RoboX, we have an interactive dialogue system in which the dialogue flow is guided by the system. The recognition technique employed when the system is acquiring the user answer is based on word spotting with a small system vocabulary. The system questions have three answer alternatives: two words corresponding to two alternative user goals (UG $= \{1,2\}$) and a third case of a undefined user goal (UG $= 0$) which can be expressed with every other word or combination of words.

6.1.1 Multimodal grounding model

The available input and output modalities on the robot platform are used in a two-phase process of multimodal grounding prior to identifying the user goals from the recognized words. The process of grounding is responsible for compensating for recognition errors that may arise due to the high noise level or uncooperative user behavior at each user dialogue turn. The grounding process monitors the four grounding states and can trigger dedicated repair actions, depending on the grounding state values. The four grounding states correspond to the binary event of *U*ser presence in *R*ange for communication (UR $= 1$) as detected using the laser modality, the binary event of *U*ser *A*ttending to the conversation (looking in the robot's camera while speaking, UA $= 1$) as detected using the video modality, the binary event accounting for *S*peech *M*odality *R*eliability (SMR $= 1$) and the event of valid *U*ser *G*oal (UG $\neq 0$). The repair actions triggered when a grounding state is not reached (e.g., UR $= 0$ or UA $= 0$) manifest themselves as sub-dialogues that may employ other modalities along with speech.

Figure 6 depicts the repair dialogue used by RoboX with the help of the two phase grounding model. The repair action dialogue sequences triggered by the grounding states are depicted in Fig. 7.

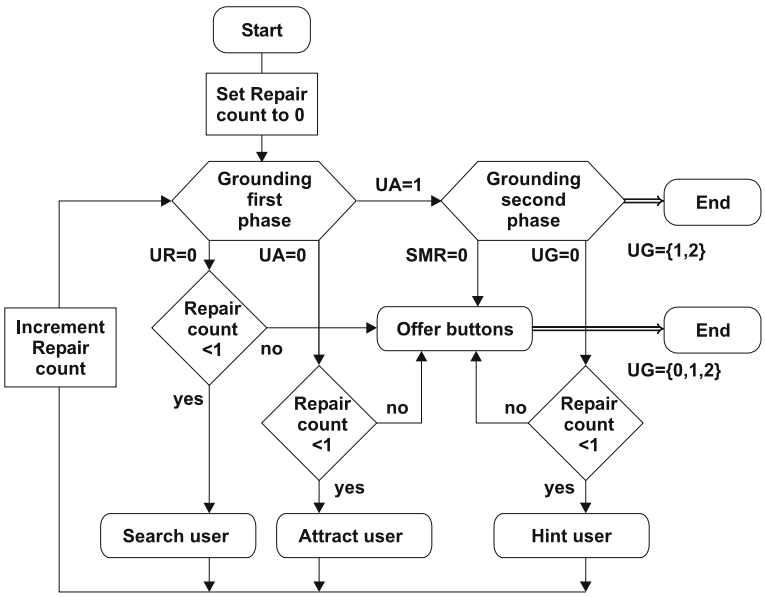


Fig. 6 Tour-guide repair dialogue

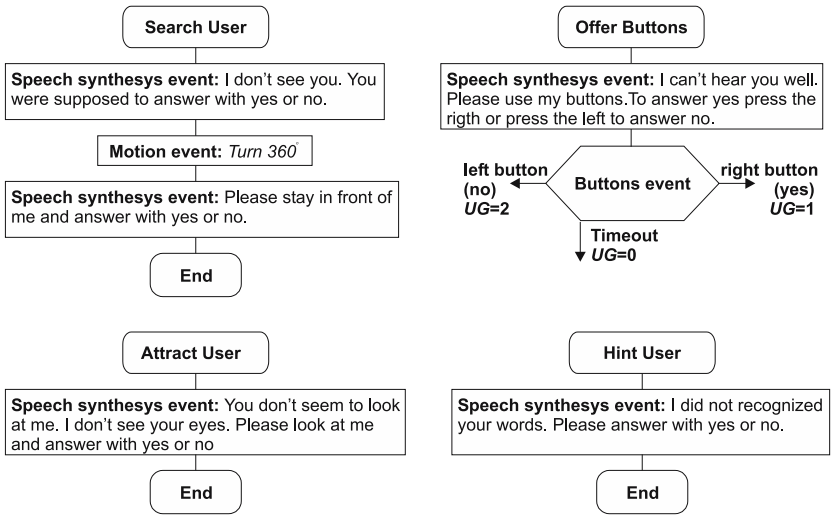


Fig. 7 Tour-guide repair action dialogue sequences

To evaluate the technical aspect of the graphical model based grounding architecture, we have to provide sufficiently large multimodal corpus for training and testing of the corresponding graphical models.

6.2 Multimodal data set collection

6.2.1 The tour-guiding scenario

In addition to the "yes/no" question/answer pair, we have added new keywords to the robot recognition vocabulary, extending the possible user goals with services related to "general lab information" or "particular lab member location". The "general lab information" tour-guide service is in fact a guided-tour in which the robot moves along the corridors of the Autonomous System Laboratory using yes/no questions to provide information about himself, the lab, the lab professors and their projects in informal interactive conversation (Table 3).

In the "particular lab member location" service the robot performs a guided-tour to the secretary's office or to the office of one of its developers. In order to distinguish between the above-specified services RoboX was configured to provide two new question/answer pairs, i.e., "tour/location" and "secretary/plamen" in his dialogue turns. The user goal mapping was preserved as in the "yes/no" case, i.e. UG = 1 corresponds to ORR = (yes|location|plamen), UG = 2 corresponds to ORR = (no|tour|secretary) and UG = 0 to ORR = GB. The particular UG interpretation depends on the dialogue turn number in the turn scenario sequence. The

Table 3 Excerpt from the *normal* tour scenario

Robot	Are you already tired? Please answer with yes or no.
User	Yes
Robot	I will show you the coffee room after we finish. I hope you will enjoy the time spent with me. Please follow me now …
Robot	Look at the poster on the wall. Do you recognize me?
User	No
Robot	Well, you know I also can't recognize myself for the moment. May be, we both should ask for a software update. Do you want me to tell you a bit more about myself?
User	Yes
Robot	Thank you, I love to speak about myself. My purpose is …

turn number was used for setting the particular speech recognition grammar as well.

We will refer to the above dialogue scenario as the *normal* tour or scenario. Our primary goal during the *normal* tour was to collect multimodal data for training and testing the grounding model in Fig. 4 as well as to observe the typical user behavior in order to identify possible communication failures that our grounding model can address in the future. Since, most of the time people were acting in a collaborative and cooperative fashion during the tour, we have created a special *simulation* tour. The goal of this tour was to provide enough communication failure examples for the training of the grounding model. During the *simulation* tour the robot himself was asking people to perform different tricky behaviors corresponding to failures at the different states of the grounding model given in Table 2. In addition to simulate noisy conditions similar to the Expo.02 exhibition conditions, each turn was replicated and noisy audio files recorded from Expo.02 were played from the robot speakers during the data acquisition process. A summary of the dialogue turns involved in the *simulation* tour scenario are given in Table 4.

To collect additional data for the speech recognizer of RoboX and to make people familiar with the robot interface, we have also designed a *tutorial* scenario. In this scenario, RoboX is explaining to people how to answer to him, asking them to repeat keywords from his recognition vocabulary several times.

6.2.2 Data sufficiency issues

A total of 60 people was involved in the data set collection experiment (20 women and 40 men). The number of people was chosen according to the standard recommendations for minimal size, speaker-independent speech corpus [10]. People were starting with the *tutorial* scenario, then they were asked to do the *normal* tour and the *simulation* tour.

During the *tutorial* scenario the new keywords to be recognized (*location*, *plamen*, *secretary* and *tour*) were repeated five times by each user. This particular number was chosen, based on the empirical recommendation that the number of training examples per recognized unit should be at least five times bigger than the number of the model parameters used in the recognition unit model. People were typically spending between 30 and 40 min communicating with the robot following the three dialogue scenarios (*tutorial*, *normal* and *simulation*). During these three dialogue scenarios, we collected data from four different input modalities of RoboX, i.e., laser, video, speech and buttons.

6.2.3 User detection

The laser modality was used for detection of the presence of a user in front of the robot (UR = 1 event, Fig. 4). The scanners were located at a height of approximately 0.5 m, which makes it possible to detect the presence of the user's legs from the scanner reading. The leg pattern typically appears as two flat minima that resemble two lines in the 1D plot of the laser scanner reading (Fig. 5c).

Whenever the user is in range for communication (within 0.5–1.5 m distance in front of the robot) the legs pattern typically appears as the closest object with respect to the the robot's front. Since we are interested in a possible user presence, the leg search is limited to the sector from the LSR (laser scanner reading) that corresponds to the robot's front. We have chosen an interval of 60° with respect to the robot front, i.e., the [150°, 210°] from the LSR (Fig. 5c). The sector width is chosen to ensure that if the user is in front of the robot within the range for communication its legs are also in this sector. When the above condition holds the flat minima produced by the user's legs have a characteristic length of the flat parts. Since these flat regions are very similar to straight lines, the flat region length corresponds to the sum of the two lines lengths. Another interesting fact is

Table 4 Dialogue turn summary for the *simulation* tour scenario

Simulation scenario
Keyword vocabulary: yes, no, location, tour, plamen, secretary

Turn number	Simulated failure	Description
1	UR = 0	User absent
2	UR = 0, NF = 1	User absent and noise
3	UA = 0	User not attending
4	UA = 0, NF = 1	User not attending and noise
5	UG = 0	User remains silent
6	UG = 0, NF = 1	User remains silent and noise
7	UG = 0	User utters out-of-vocabulary (OOV) words
8	UG = 0, NF = 1	User utters OOV words and noise
9–14	UG ≠ 0, NF = 1	User utters each vocabulary keyword in noise

that these two "lines" appear parallel to the x-axis into the 1D plot of the LSR. Since the robot is moving alongside a corridor such parallel patterns appear very rarely in the case of a missing user or they will be quite far from the robot. On the other hand, a histogram of the LSR produces high valued bins whenever such parallel structures are observed in the signal. The number of bins has to be chosen with respect to the needed precision when legs are detected. We chose 45 bins that divide the range of the SICK scanner into equally spaced intervals of 20 cm. In the case of a user present in front of the robot the first histogram bin is significantly higher compared to the case of no object, given that the robot is always looking alongside the corridor. Therefore, we have chosen the first bin value for the continuous LSR variable used by the Bayesian network in Fig. 4a.

6.2.4 User face detection

The video modality was used for detecting a user attending to the conversation (UA = 1, Fig. 4). Given the presence of a user, the robot has to detect if the user is attending to the conversation. We assume that presence of a user's frontal face in the video frames for an interval of time of at least 0.8 s is sufficient to ensure that the user is attending while providing her/his spoken answer. The video stream is providing 25 frames per second on the average (Fig. 5). In order to provide evidence for the state of the UA variable from Fig. 4 we use a face detector based on the modified algorithm of Viola and Jones [27,40]. To detect the user as attending we look for the binary event of face detected into 10 consecutive frames in the video stream. We assign this observed event a binary variable FD (face detected) and we use it in the Bayesian network in Fig. 4a.

6.2.5 Speech modality reliability

The speech modality is used to obtain values for the observed variables in the Bayesian network in Fig. 4b. The speech recognition system provides the values for the observed recognition result — ORR variable for each user turn in dialogue that are subsequently interpreted into IRR (interpreted recognition result into user goals) values. Each robot dialogue turn contains a question offering two possible services. The answer of the user is mapped into three possible user goals UG = 1 − first possible service, UG = 2 − second service and UG = 0 − undefined user goal at each dialogue state.

To measure the acoustical conditions affecting the noise factor (NF) we use a signal-to-noise ratio (SNR) related measure. The SNR can be defined as the ratio of the average energy of the speech signal divided by the average energy of the acoustic noise in dB. As in our case we have a single channel speech signal we estimate these energies based on two passes of audio signal acquisition. The first pass is just before the final question of RoboX and is 0.5 s long. The second pass is during the user answer and is limited to 2 s which was estimated to be a sufficient duration given the keyword vocabulary of RoboX. The signal n acquired in the first pass is associated with noise, while the signal s from the second pass is associated with speech. Our SNR-related modality quality measure (QM) is given by the formula:

$$\mathrm{QM} = 10 \log_{10} \frac{\frac{1}{N}\sum_{i=1}^{N} s^2(i)}{\frac{1}{M}\sum_{i=1}^{M} n^2(i)}, \tag{8}$$

where $\{s(i)\}, i = 1, \ldots, N$ is the acquired speech signal containing N samples, and $\{n(i)\}, i = 1, \ldots, M$ is the acquired noise signal containing M samples. As the audio input of RoboX is sampled at $f_s = 16$ kHz, then $N = 32,000$, and $M = 8,000$.

6.2.6 Database organization

The buttons modality of RoboX was used during the data collection to auto-assign user goals to the spoken answers of the user during the *normal* tour. In that case the users were asked to press one of the four buttons of RoboX corresponding to their spoken answer. The buttons status was recorded during the phase of input

modality data acquisition, however the actual decision for the next robot dialogue turn was based solely on the speech recognition result (ORR) during the interaction with the user. In the remaining two scenarios (*tutorial* and *simulation*) the user goals (*UG* values) were *a − priori* known from the designing stage. The use of UG predefined scenarios (*tutorial* and *simulation*) and the buttons modality permitted automatic data tagging for all of the unobserved variables (UR, UA, NF, SMR, UG) in the robot grounding model. UG was set to 0, whenever UR or UA were 0. The NF values were set to 1 during the "noisy" turns in the *simulation* scenario (see Table 4) and 0 otherwise. According to its definition, SMR is 1 when UG coincides with IRR and is 0 otherwise.

6.3 Technical evaluation experiments

6.3.1 Component level evaluation

In the component level evaluation of the multimodal grounding we asses the accuracies of the grounding state predictors as well as the accuracy of the final user goal identification. The accuracies are calculated for the baseline tour-guide dialogue system and compared with an alternative system. The alternative system employs grounding and argmax criteria on each of the grounding states posteriors to select a state value. It is named the "Argmax BN" system.

In the component level evaluation we adopt an accuracy metric similar to the word recognition accuracy. In our case, the recognition task is to detect a keyword (e.g., yes, no, location, etc.) or a "garbage" word (GB) in the spoken input. Therefore, the errors can be only of substitution type and we can directly evaluate the user goal accuracy using the formula:

$$\mathrm{Acc} = 100\left(1 - \frac{N_S}{N}\right), \tag{9}$$

where N_S is the number of substitutions, and N is the total number of testing examples. The same formula is used in the case of evaluating the grounding state prediction accuracy.

6.3.2 Accuracy of the "Argmax BN" system versus baseline system

The collected data set was used to train and test the grounding model networks. The full data set was used for training and testing of the attendance phase Bayesian network. Given the two phase grounding model of RoboX, the speech reliability Bayesian network was used only after detecting the event UA = 1 (User Attending) in the first phase of grounding. Hence, in the training of the second phase network, we do not really need data from the records for which UA is zero. Such data will very rarely appear in the second phase of grounding. For that reason the speech reliability phase Bayesian network was trained and tested on a partition of the full data set containing "clean" recordings (NF = 0) from the *tutorial* scenario and "noisy" ones (NF = 1) from the *simulation* scenario.

To test the accuracies of the individual grounding state predictor variables UR, UA and SMR we have run 50 cross-validation tests. Training and testing portions were chosen from the full and the partitioned data set each time at random. The size of the training portion was two times bigger than the testing portion. Values for the posteriors $P(\mathrm{UR}|E_1)$, $P(\mathrm{UA}|E_1)$ from the Attendance BN (Fig. 4) and $P(\mathrm{SMR}|E_2)$ from the speech Reliability phase BN were calculated for each testing sample ($E_1 = \{\mathrm{LSR}, \mathrm{FD}\}$ in the first case and $E_2 = \{\mathrm{IRR}, \mathrm{SNR}\}$ in the second case).

The values for the corresponding state predictor variables were assigned using the argmax criteria on the corresponding posterior probabilities. The tests were done for the events UR = 1, UA = 1, SMR = 1 computing corresponding accuracies. We have also done the tests for the noise factor event, i.e., NF = 1. The accuracies are calculated as the number of correct classifications minus the number of substitutions divided by the number of examples per class. The total number of training and testing examples were 1,900 and 949 for the first phase of grounding and 1,404 and 701 for the second phase. The accuracy statistics are given in Table 5.

To test the efficiency of the two phase grounding model in detecting the recognition errors, we have done the following experiment: we have trained the Bayesian networks in Fig. 4 with a single iteration of the cross-validation test. The testing examples were provided first to the Bayesian network for the first grounding phase. If UA = 0 was calculated to hold after applying the argmax criterion the user goal was set to UG = 0. Otherwise, the examples were provided to the second grounding phase Bayesian network. After computing the posterior distribution $P(\mathrm{SMR}|E_2)$, if SMR = 1 was true, the IRR result (the user goal based on the speech recognition only) was used to assign a user goal. Otherwise, if SMR = 0 was selected after applying the argmax criterion, we were setting the UG to its tagged value from the testing data. We assume that if the speech modality is unreliable and the user is requested to use the buttons the user goal is normally provided without errors.

Table 5 50 cross validation accuracy statistics for user attendance and speech reliability grounding phase BN models

Attendance BN Acc stats with 1,900/949 train/test samples			
Acc UR %	UR = 1	UR = 0	Total Acc
μ	98.1	100	99.1
σ	0.3	0	0.3
Acc UA %	UA = 1	UA = 0	Total Acc
μ	94.3	90.7	94.0
σ	0.6	3.2	0.6
Reliability BN Acc stats with 1,404/701 train/test samples			
Acc SMR %	SMR = 1	SMR = 0	Total Acc
μ	89.9	67.6	83.5
σ	0.9	2.8	1.1
Acc NF %	NF = 1	NF = 0	Total Acc
μ	80.6	93.5	90.6
σ	3.1	0.9	0.8

The accuracy of IRR (the user goal based on the speech recognition only) was calculated and compared with that of UG after applying the two grounding phases. The results are presented in Table 6.

As can be seen from Table 5, the grounding state predictors function significantly above chance level. Thus, should the grounding level need to be assessed, the cause of the communication failure can be located and remedied. This statement seems to be strongly supported by the results from our evaluation experiment as well. As can be seen from Table 6 the use of the Bayesian networks in Fig. 4 for the two phases of grounding has resulted in a significant improvement in the accuracy of the user goal identification. The gain in performance is due to the improved identification of the garbage case UG = 0, which in turn is due to the good detection rate of the event UA = 1 in the first phase of grounding when using the Bayesian network in Fig. 4a. Modeling of the event of error in user goal identification based only on the observed speech recognition results in the second phase of grounding and the availability of an alternative input modality (interactive buttons) can enable even further improvement in the user goal identification as demonstrated in Table 6.

6.4 System-level evaluation

Mobile service robots in general and tour-guide robots in particular are physical agents that act in the real world, sensing changes in the environment through their input modalities (e.g., speech) and performing actions through the output modalities (e.g., synthesized speech). The performed action at each time, given the information acquired from the input modalities at that time has to be chosen in order to maximize a performance metric. The performance metrics are measurable quantities related to success criteria that evaluate how successful the agent is in fulfilling its communicative tasks. For the tour-guide communicative tasks we adopt the following success criteria: a tour guide robot is considered successful in its interaction with the user if:

Criterion 1 The user is attending to the conversation, which means that the states in the first phase of grounding are reached in all initiative/response pairs, during one full tour-guide dialogue scenario. In this way, we ensure that information is successfully conveyed to the user.

Criterion 2 The user choice is considered after each user turn in dialogue. In other words, user goals are correctly identified during the dialogue. This additionally requires that the states in the second phase of grounding are reached in all initiative/response pairs, during dialogue.

The criteria are ordered according to their decreasing significance for the usability perspective of the voice-enabled tour-guide robot. If a user is always present and attending (Criterion 1) in front of the robot, we assume that the level of user interest and interface usability is high. Although user goal identification accuracy is important from the perspective of the tour-guide ability to provide desired service it is not assumed to be more important than the ability of the the tour-guide to attract its users. The final goal of providing specific information should not contradict the goal of keeping the user involved and informed according to her/his intent.

Criterion 1 can be quantified by the parameter "user attendance rate". We define it as equal to the number of times during the dialogue that the user was attending to the conversation (UR = 1 and UA = 1) divided by the total number of robot dialogue turns:

Table 6 Statistics about user goal identification before (IRR) and after grounding (UG)

Total Acc IRR	IRR = 0	IRR = 1	IRR = 2
67.1 %	63.3 %	65.3 %	72.8 %
Total Acc UG	UG = 0	UG = 1	UG = 2
90.2 %	95.0 %	84.4 %	91.2 %

$$\mathrm{UAR} = \frac{1}{N}\sum_{t=1}^{N} I_t(\mathrm{UA}=1), \tag{10}$$

where UAR is the user attendance rate and $I_t(\mathrm{UA}=1)$ is the indicator function of the event UA = 1 at each dialogue turn $t = \{1,\ldots,N\}$. The number of dialogue turns N in the definition does not include the additional repair turns. In order to have a "fair" measure, the indicator function has to be used with a priori annotated reference state after looking at the collected dataset.

Criterion 1 can be also related to the dialogue task success metric. In the case of tour-guiding, completing a full scenario with a user attending to the conversation can be seen as a successful task completion.

Criterion 2 can be directly quantified by the user goal identification accuracy during the spoken interaction. In addition, to evaluate the efficiency of considering the user choice using the grounding model for multimodal dialogue repair we introduce the so-called *Repair proportion*. The repair proportion is closely related to the reported turn repair ratio metric in dialogue system evaluation [9]. The *Repair proportion* is calculated with respect to the number of robot dialogue turns in the dialogue, i.e.,

$$\mathrm{RP} = \frac{N_{\mathrm{repairs}}}{N}, \tag{11}$$

where RP denotes the repair proportion measure, N_{repairs} corresponds to the total number of repair turns, and N corresponds to the number of dialogue turns in the current dialogue scenario as in Eq. 10.

All the metrics specified above had to be calculated for the baseline dialogue system and after performing grounding and corresponding repairs to evaluate the yield from applying the error repair techniques, using the *normal* tour scenario. However, the data collected with the *normal* tour for the purpose of the component-level technical evaluation (Sect. 6.2) were recorded under controlled user conditions. In order to get the real figures using the above system-level metrics, we need an interactive scenario that is close to the real conditions of application. For that purpose we have performed a subjective user satisfaction test, where the system-level objective metrics are calculated and compared with results from user surveys on the interactive system usability.

6.5 Subjective user satisfaction tests

In the subjective user test 22 users (7 female/15 male) are asked to perform the *normal* tour scenario. They were not given any additional information apart from a very general description of the robot and its input modalities. In addition, the tour itself is initiated with a short help on how to communicate with the robot. During the tour the user was advised to behave as natural as possible. The user was not obliged to follow the whole presentation if she/he gets very bored or for any other reason was willing to abandon the robot. Table 7 depicts statistics about the people involved in the experiment.

The main focus of the experiment was on the ability of the robot to keep its user involved and attending to the interaction. At the end the user was given to fill in a survey that aims at assessing the user satisfaction with the interactive performance of the RoboX system.

During the user satisfaction test the multimodal user input (speech/video/laser) was recorded along with the status of the repair dialogue sequence. This status includes the number and type of the performed repairs during the repair dialogue sequence, including the detected grounding state value. At the end of each *normal* tour the real grounding state values manually annotated are compared with the automatically detected ones and system-level evaluation metrics are calculated. To evaluate the gain from the use of repair actions, the system-level evaluation metrics are calculated before and after the repair sequence. The results after calculating the system-level evaluation metrics are presented in Table 8.

The two subjective measures of system usability presented in Table 8 (Dialogue quality and Recognition performance) were extracted from the user answers to questions 1 and 7 in the survey. These questions along with the answers statistics from the 22 participants are depicted in Fig. 8. The user satisfaction with the repair sequence performance is depicted similarly in the same figure.

7 Dicussion

7.1 Efficiency of the repair strategy

Introducing two phases of grounding has added the advantage that we do not need to provide all the evidence from the input modalities in the first grounding

Table 7 Personal information about the user satisfaction test participants

User No	Occupation	Sex	Age between		English Speaker	Familiarity with Robots		
						1	2	3
1	Ph.D. student	Female	25	35	Non	No	No	No
2	Student	Female	25	35	Non	No	No	No
3	Student	Female	25	35	Non	No	No	No
4	Unemployed	Male	36	45	Non	No	No	Yes
5	Ph.D. student	Male	25	35	Non	Yes	Yes	Yes
6	Assistant	Male	25	35	Non	Yes	No	Yes
7	Ph.D. student	Male	25	35	Non	No	No	No
8	Ph.D. student	Male	18	24	Non	No	No	No
9	Ph.D. student	Male	18	24	Non	Yes	No	No
10	Post-doc	Male	36	45	Non	No	No	Yes
11	Professor	Male	25	35	Non	No	No	Yes
12	Ph.D. student	Female	25	35	Non	Yes	Yes	No
13	Assistant	Male	25	35	Non	Yes	Yes	Yes
14	Ph.D. student	Male	25	35	Non	No	No	No
15	Ph.D. student	Female	25	35	Non	No	No	No
16	Ph.D. student	Male	25	35	Non	No	No	No
17	Ph.D. student	Male	25	35	Non	No	No	No
18	Ph.D. student	Male	25	35	Non	No	No	Yes
19	Ph.D. student	Male	25	35	Non	No	Yes	Yes
20	Ph.D. student	Female	25	35	Non	No	No	No
21	Musician	Female	25	35	Non	No	No	No
22	Scientist	Male	46	55	Non	Yes	Yes	Yes
Average	Ph.D. student	68%	26	36	100%	73%	77%	59%
Comment	Mostly	Male	–	–	Non	No	No	No

1 Have you ever used a real robot?
2 Controlled a robot with voice?
3 Used speech recognition software?

Table 8 Results for the system level evaluation metrics before and after grounding

Users: 22	Task success	Repair proportion	UG Acc[a] before repair	UG Acc after repair	UAR[b] before repair	UAR after repair
Average	0.91	0.62	0.65	0.94	0.89	0.95

[a] UG Acc User Goal Accuracy
[b] UAR User Attendance Rate

phase when the robot is concerned with the issue of user presence and attention to the dialogue. Running a speech recognition process at this stage will just result in unnecessary workload for the robot system. On the other hand the task of people detection and face detection are also required for the purpose of safe navigation and situation awareness of the mobile robot. They are typically implemented and running all the time and their status is already available. Thus, the two phase separation of the grounding process contributes to the efficient utilization of the robot modality information. It also defines an efficient strategy for communication failure detection and repair. Given the dependencies in the Bayesian network in Fig. 4a inferring that UA = 1 is causally related with UR = 1. In other words presence of a face in the video stream would mean presence of legs in the laser scanner reading. Thus, the state of user presence (UR variable) is checked only when UA is inferred to be 0, using the argmax criterion on the UA posterior probabilities.

7.2 Grounding with multimodal dialogue repairs

In order to consider a grounding state as being reached, the robot seeks for a probability above chance level for a particular value (e.g., UR = 1) of the modality event associated with that state given the evidence from its input modalities. Hence, we have established a grounding criterion for the purpose of service robot dialogue that is based on the probability of the modality events associated with the grounding states in a two phase grounding model.

Whenever a failure to reach a state is detected the multimodal grounding model can be used to trigger multimodal dialogue repair actions (grounding actions). For example, failure to reach grounding state S0 (UR = 0)

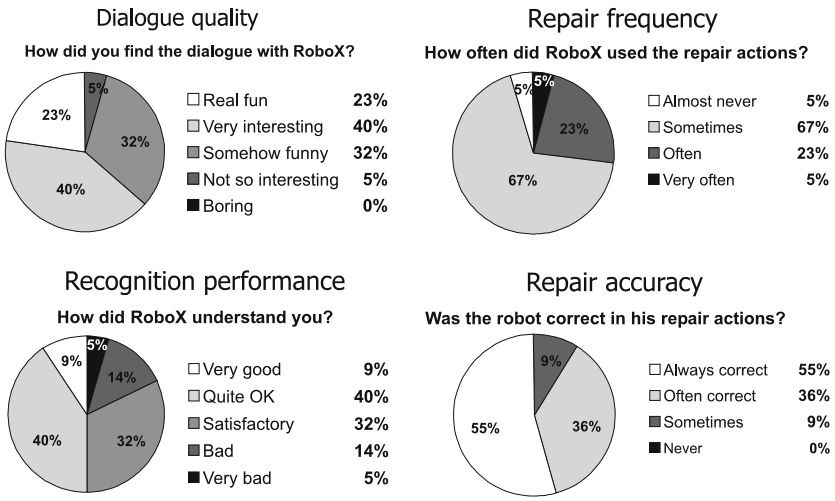


Fig. 8 User satisfaction with the dialogue quality, the recognition performance, the repair frequency and accuracy during dialogue

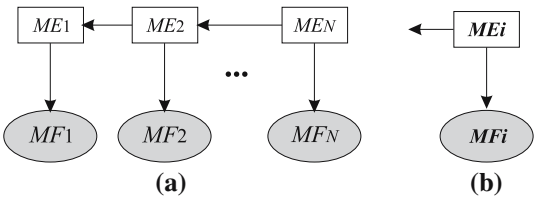


Fig. 9 Bayesian network for grounding. (**a**) Slice related to a modality event and its feature, (**b**) full topology

can trigger a dialogue repair action dedicated to finding a user ("Search visitor", Fig. 7). This repair action can combine speech synthesis as well as the move modality of the robot in the process of user search. If the second state of grounding is not reached (UA = 0), speech as well as the robot expressive face can be used to attract the attention of the user. Buttons can be used as an alternative input when the grounding state S2 is not reached (SMR = 0). At the end, if the user goal is still undefined (UG = 0) the expressive face along with the speech synthesis can be used to hint the user for the possible keywords that her/his answer can contain. In order to model the robot preferences on a particular repair action the framework of decision networks and utilities [34] can be directly used with the presented model of grounding. The grounding model can be also applied in more complex dialogue systems employing keyword spotting as well as continuous recognition systems in a system-initiative or mixed-initiative dialogue setting. In particular, the first phase of grounding would not require any modification or changes in the network topology. As long as we preserve the user goal-oriented turn structure of the service dialogue, the second phase of grounding may not require any changes in the network topology either. We have to mention however that depending on the representation of the user goal [12] and the type of the recognition task involved (keywords, continuous speech), the grounding model in its second phase may need additional states associated with speech-based dialogue repair acts well known from the spoken dialogue literature (e.g., different kinds of confirmation and disambiguation grounding acts [3]).

7.3 Scalability of the grounding model

Extending the model with additional modalities and user goals should be done after taking into account the complexity issues concerning the framework of Bayesian networks. The computational complexity of exact inference in Bayesian networks with conditional Gaussian pdfs is NP hard [7,28]. In our case, however, the use of two phases of grounding and special Bayesian network topologies lead to great reduction in the computational demands for inference in the Bayesian networks in Fig. 4. In addition, the continuous variables are all observed, which avoids the problem of marginalizing continuous variables.

The Bayesian network in the first phase of grounding is a member of a special class of Bayesian network topologies: the polytree or the singly-connected networks that allow linear dependence of the number of computations needed by exact inference on the size of the network [28]. A polytree network is a Bayesian network in which there is only one path between any two variables. The Baysian network in Fig. 4a is a polytree network that is a subtree from a more general topology depicted in Fig. 9a. This network is composed of slices corresponding to distinct modalities. Each such slice contains a modality event causally-related to a modality feature (Fig. 9b). The full topology in Fig. 9a can model the causal chains similar to the one in the first phase of grounding. For example, the modality event ME_2 can

be the UA event (User is attending to the conversation) that in turn is seen as the cause for next modality event ME_1 = UR (User is present staying in close range in front of the robot). In that case we end up with the network in Fig. 4a. The incorporation of a new modality and its event/feature is straightforward — we just add a new slice in the causal chain. For example, ME_3 can be an event related to the the event of a user who is speaking. This event can be seen as cause behind a feature MF_3 given by a voice activity detector. A slice can also represent another event from the same modality. For example, the event of a speaking user can be related to a video modality feature related to detecting movements of the user's lips.

Using exact inference algorithms like the junction tree algorithm or variable elimination ([28], Appendix B) with the above BN topology will result in linear computational complexity $O(N)$ with the number N of the involved modalities events.

In the second grounding phase the definition of the SMR event also allows reduction in the number of operations needed by inference in the corresponding Bayesian network as already discussed in Sect. 5.2 (Eq. 7). These observations demonstrate the important fact that particular Bayesian network topologies offered for multimodal grounding offer substantial reduction in the computational complexity of inference.

7.4 System-level evaluation metrics and system usability

Despite the limited available data (22 participants) the results from the user-based evaluation also supported the fact that the proposed grounding model can contribute to a significant gain in the accuracy of final (after the repairs if any) user goal identification (Table 8), as well as a gain in user attendance rate. Hence, the use of multimodal grounding can enhance the usability of the service robot interactive system. The above statement is also supported by the high average task success and UAR (user attendance rate) with the 22 users (Table 8). It has to be taken into account that in real application conditions users may be less cooperative than the participants in the presented user study.

According to the subjective usability measures (Dialogue quality and Recognition performance, Fig. 8) and the technical measures from Table 8, we can conclude that the RoboX dialogue scenario was appealing to the user and that the robot was efficient in providing its information to its user. In order to provide finer interpretation and motivation behind the above statement in the following section we perform a communication failure analysis of the logged grounding state values during the user tests. The user feedback is also analyzed to provide guidelines for further improvement of the interactive system of RoboX.

7.5 Communication failure analysis

During the user tests there were two cases in which the interaction between RoboX and its user has resulted in a communication failure (the robot was unable to identify a valid user goal after two consecutive repair actions). In the first case, the user wanted to experiment with the robot on purpose, and did not answer the robot's questions to see what will happen. After the buttons repair timed out RoboX left, informing that if the user is still there, they can meet again near the coffee room. The second case was due to technical problems with the video camera. As a result the user was repeatedly asked to look at the robot in the eyes without a real reason for such a repair action during several consecutive system turns. The increased repair activity frustrated the user, who finally left the robot to look for a human operator. As a result RoboX moved to the coffee room area, where after re-plugging the camera cable, the robot operated without any further technical problems.

Among the main sources for errors in user goal identification, when only speech recognition was used, were the background noise, particular user accents or clipping of the user answer because of the two seconds acquisition time interval. In such conditions the subsequent repairs were useful giving the robot a second chance for input acquisition, as well as the alternative to use buttons in the case of noise and in the second repair pass. Due to the "two phase" SNR calculation technique described in Sect. 6.2.5, whenever the user answer was preceded by non-stationary (temporary) burst of noise, the robot was declaring the user answer as very noisy, although it was actually recorded in clean audio conditions. Such audio disturbances could potentially result in wrong repair actions related to speech modality reliability.

Detecting the state of user attendance depends directly on the frontal face detection accuracy. With proper user positioning with respect to the camera the errors in face detection were mainly due to adverse illumination conditions, i.e., sun flare from behind the user during the day or insufficient light in the evenings. The other main source of errors resulted from the user posture or camera adjustment. In these cases, typically, part of the face was remaining outside the visual range of the camera. This was often the case with users that tended to stay too close to the camera or tended to

bend toward the microphone while answering. Clipped faces also resulted with the users that were staying aside instead of directly facing the robot's front. In the last case, the "Attract user" repair was particularly useful for successful grounding (reaching the state of UA = 1).

Finally, in some repair sequences users pressed a wrong button that resulted in a wrong user goal assignment. Nevertheless, in general, users remained interested in the conversation. Sometimes, incorrect user goals remained even unnoticed or were attributed to the humoristic character of the robot.

7.6 User feedback

As seen from Fig. 8 most of the user test participants described the interaction with RoboX as funny and entertaining. Since many of them were unfamiliar with robots (Table 7) and with the dialogue scenario, the system driven dialogue did not make a bad impression on them. There were no recommendations in the survey that explicitly suggested changing the dialogue initiative. However, several persons recommended the robot to use more keywords and be more personal with them (e.g., asking for their name and using this information in the scenario). One of the users even started answering spontaneously with natural speech, but after the second question he understood that the robot preferred keywords, and adjusted his spoken answers appropriately.

People found the humoristic style of the tour guide as appropriate for its task. When asked if they would prefer "more serious" tour-guide, all users answered negatively. The positive attitude towards communicating with the robot did not change even when the robot's speech recognizer was not performing well all the time. However, in these cases the repair style was found to be important in order to avoid the impression that the system does not perform well. One user that exhibited low recognition performance (numerous "Hint user" repairs in more than two consecutive dialogue turns) recommended that the input modality should be permanently switched to buttons after given number of repairs related to speech recognition. Another user perceived the repeating "Attract user" repair as impolite, suggesting that the repair text should vary to overcome this impression. In two of the cases in the study with high repair activity (Repair proportion > 1), the users reported that their high concentration in answering the robot has distracted them from the normal process of listening to the information content provided by the robot during the tour-guiding scenario. However, most of the users (86 % — 19 out of 22 people) reported that the repair actions helped them stay involved and more interested in the dialogue. The repair actions seemed to distract people from their sometimes "destructive" desire to investigate and experiment with how they can put the robot in difficulty. We have to mention however that users were mostly highly educated people aware of the fact that the robot is recording their activities. Throughout the scenario the user preference toward the two alternative input modalities remained mostly in favor of speech and the combined use of speech and buttons. Only two of the users preferred permanently the use of buttons.

8 Conclusion

In this paper, we have introduced a multimodal state-based model for low-level grounding in conversation with a service robot under noisy acoustic conditions. The model was motivated by reducing the risk of communication failures due to incorrect user goal identification with unprepared users in typical noisy robot deployment conditions. The model exploits the multiple modalities available in the service robot system to provide evidence for reaching grounding states. In order to handle the speech input as sufficiently grounded (correctly understood) by the robot, four proposed states have to be reached in two distinct phases of grounding.

The initial two states in the first grounding phase are related to the events of presence of a user who is attending to the robot conversation (looking at the robot). A Bayesian networks combining information from the laser and video modality was used to estimate probabilities that the grounding states have been reached. The remaining two states in the second phase of grounding were related to the state of reliable speech modality and the state of valid user goal, i.e., a user goal that can be mapped into a service provided by the robot. The speech modality reliability was explicitly modeled by the event of error in the user goal identification based on the observed recognition result. Another Bayesian network was used to model the dependencies between the event of speech modality reliability, the user goal and the speech recognition result as well as signal–domain measure related to the level of acoustic noise. The criterion used to consider the conversation as grounded at each particular grounding state was based on the probability of the grounding state-related events, estimated by the Bayesian network.

The use of Bayesian networks enabled explicit modeling of the uncertainties intrinsic to speech and other input modalities' information during human–robot interaction, using an intuitive graph-based probabilistic

framework. The use of two distinct phases of grounding and special topologies in the Bayesian networks resulted in reduced number of computations needed for probabilistic inference, contributing to the system scalability and modularity. In particular, using a polytree (singly-connected) BN topology in the first grounding phase has allowed reduction from exponential to linear number of operations in the number of used modalities needed by inference. At the same time the two-phase grounding model reduces the workload for the speech recognizer, as speech recognition is performed only in the second grounding phase.

The performance of the model was tested with real data from a database, collected during the operation of the service robot RoboX as a tour-guide in the Autonomous System Laboratory at EPFL. The evaluation was done in both objective (employing technical performance metrics) and subjective (user-based) evaluation experiments.

Both technical and subjective user satisfaction evaluation supported the fact that the proposed grounding model can contribute to a significant gain in the accuracy of the final user goal identification, as well as a gain in user attendance rate. The evaluation showed that generally, the use of Bayesian networks for multimodal low-level grounding enhances the usability of the service robot voice-enabled communication interface.

Acknowledgments

The authors would like to thank Professor Roland Siegwart and his team at the Autonomous Systems Laboratory, Swiss Federal Institute of Technology, Lausanne (EPFL) for their support and participation in the experiments presented in the paper.

References

1. Aji SM, McEliece RJ (2000) The generalized distributive law. IEEE Trans Inf Theory 46(2):325–343
2. Aoyama K, Shimomura H (2005) Real world speech interaction with a humanoid robot on a layered robot behavior architecture. In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, ICRA05, Barcelona, Spain, pp 3825–3830
3. Brennan SE, Hulteen EA (1995) Interaction and feedback in a spoken language system: a theoretical framework. Knowl Based Syst 8(2–3):143–151
4. Burgard W, Cremers AB, Fox D, Hhnel D, Lakemeyer G, Schulz D, Steiner W, Thrun S (1999) Experiences with an interactive museum tour-guide robot. Artif Intell 114(1–2): 1–53
5. Clark H, Brennan S (1991) Perspectives on socially shared cognition Grounding in Communication American Psychological Association, Washington, pp 127–149
6. Clark HH, Schaefer EF (1989) Contributing to discourse. Cognit Sci 13(2):259–294
7. Cooper GF (1990) The computational complexity of probabilistic inference using Bayesian belief networks (research note). Artif Intell 42(2–3):393–405
8. Drygajlo A, Prodanov P, Ramel G, Messier M, Siegwart R (2003) On developing voice enabled interface for interactive tour-guide robots. Adv Robot 17(7):599–616
9. Dybkjaer L, Bernsen NO, Minker W (2004) Evaluation and usability of multimodal spoken language dialogue systems. Speech Communi 43(1–2):33–54
10. Gibbon D, Moore R, R. Winski, e. (1997) Handbook of standards and resources for spoken language systems. Mouton de Gruyter, Berlin
11. Gibbon D, Mertins I, R. Moore, e. (2000) Handbook of multimodal and spoken dialogue systems: resources, termonology and product evaluation. Kluwer, Dordchet
12. Hong J-H, Song Y-S, Cho S-B (2005) A hierarchical bayesian network for mixed-initiative human-robot interaction. In: 2005 IEEE International Conference on Robotics and Automation, ICRA 2005 Barcelona, Spain, pp 3819–3824
13. Horvitz E, Paek T (1999) A computational architecture for conversation. In: UM '99: Proceedings of the seventh international conference on User modeling, Springer, New York, Secaucus, NJ, USA, pp 201–210
14. Horvitz E, Paek T (2000) Deeplistener: Harnessing expected utility to guide clarification dialog in spoken language systems. In: ICSLP 2000: 6th international conference on spoken language processing, Beijing, China
15. Horvitz E, Paek T (2001) Harnessing models of users' goals to mediate clarification dialog in spoken language systems. In: UM '01: Proceedings of the 8th international conference on user modeling 2001, Springer, Berlin, London, UK, pp 3–13
16. Huang X, Acero A, Hon H-W (2001) Spoken language Processing: a guide to theory, algorithm and system development, 1st edn. Prentice Hall
17. Huttenrauch H, Green A, Norman M, Oestreicher L, Eklundh K (2004) Involving users in the design of a mobile office robot. IEEE Trans Syst Man Cybern, C 34(2):113–124
18. Jensen B, Froidevaux G, Greppin X, Lorotte A, Mayor L, Meisser M, Ramel G, Siegwart R (2002a) The interactive autonomous mobile system robox. In: International Conference on intelligent robots and systems, IROS 2002, Lausanne, Switzerland, pp 1221–1227
19. Jensen B, Froidevaux G, Greppin X, Lorotte A, Mayor L, Meisser M, Ramel G, Siegwart R (2002b) Visitor flow management using human-robot interaction at expo.02. In: Workshop: robotics in exhibitions, IROS 2002, Lausanne, Switzerland
20. Jensen B, Tomatis N, Mayor L, Drygajlo A, Siegwart R (2005) Robots meet humans—interaction in public spaces. IEEE Trans Ind Electron 52(6):1530–1546
21. Jensen F (1996) An introduction to Bayesian networks, 1st edn. UCL Press
22. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) An introduction to variational methods for graphical models. Machine Learn 37(2):183–233
23. Josifovski L (2002) Robust automatic speech recognition with missing and unreliable data. Ph.D. thesis, Department of Computer Science, University of Sheffield, UK
24. Kleinehagenbrock M, Lang S, Fritsch J, Lömker F, Fink GA, Sagerer G (2002) Person tracking with a mobile robot based on multi-modal anchoring. In: Proceedings IEEE International workshop on robot and human interactive communication (ROMAN), IEEE Berlin, Germany, IEEE, pp 423–429

25. Lang S, Kleinehagenbrock M, Hohenner S, Fritsch J, Fink GA, Sagerer G (2003) Providing the basis for human–robot-interaction: a multi-modal attention system for a mobile robot. In: ICMI '03: Proceedings of the 5th international conference on multimodal interfaces, NY, USA ACM Press, New York, pp 28–35
26. Li S, Haasch A, Wrede B, Fritsch J, Sagerer G (2005) Human-style interaction with a robot for cooperative learning of scene objects. In: ICMI '05: Proceedings of the 7th international conference on Multimodal interfaces, ACM Press, New York, NY, USA, pp 151–158
27. Lienhart R, Maydt J (2002) An extended set of haar-like features for rapid objection detection. IEEE ICIP, 900–903
28. Murphy K (2002) Dynamic bayesian networks: representation, inference and learning. Ph.D. thesis, U.C. Berkeley
29. Nakadai K, Hidai K, Mizoguchi H, Okuno HG, Kitano H (2001) Real-time auditory and visual multiple-object tracking for humanoids. In: Proceedings of the 17th international joint conference on artificial intelligence, IJCAI 2001, Seattle, Washington, USA, pp 1425–1436
30. Paek T, Horvitz E (1999) Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational systems. In: Brennan SE, Giboin A, Traum D (eds) Working papers of the AAAI fall symposium on psychological models of communication in collaborative systems, American Association for Artificial Intelligence, Menlo Park, California, pp 85–92
31. Paek T, Horvitz E, Ringger E (2000) Continuous listening for unconstrained spoken dialog. In: ICSLP 2000: 6th international conference on Spoken Language Processing, Beijing, China
32. Pavlovic VI (1999) Dynamic Bayesian networks for information fusion with application to human-computer interfaces. Ph.D. thesis, University of Illinois Urbana-Champaign
33. Prodanov P, Drygajlo A (2005a) Bayesian networks based multi-modality fusion for error handling in human-robot dialogues under noisy conditions. Speech Communi 45(3): 231–248
34. Prodanov P, Drygajlo A (2005b) Decision networks for repair strategies in speech-based interaction with mobile tour-guide robots. In: Proceedings of international conference on robotics and automation, IEEE ICRA 2005, Barcelona, Spain
35. Russell S, Norvig P (2003) Artificial intelligence: a modern approach. 2nd edn. Prentice Hall
36. Sidner CL, Kidd C, Lee C, Lesh N (2004) Where to look: A study of human-robot engagement. In: Proceedings intelligent user interfaces (IUI), Funchal, Island of Madeira, Portugal, pp 78–84
37. Tasaki T, Komatani K, Ogata T, Okuno HG (2005) Spatially mapping of friendliness for human-robot interaction. In: Proceedings of IEEE/RSJ international conference on intelligent robots and systems (IROS 2005), Edmonton, Alberta, Canada
38. Traum D (1999) Computational models of grounding in collaborative systems. In: AAAI fall symposium on psychological models of communication, pp 124–131
39. Traum DR, Dillenbourg P (1998) Towards a normative model of grounding in collaboration.
40. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: IEEE computer society conference on computer vision and pattern recognition (CVPR), ISSN: 1063-6919, vol 1, pp 511–518