Graphical Models for Dialogue Repair in Multimodal Interaction with Service Robots

P. Prodanov[1], J. Richiardi[2] and A. Drygajlo[2]

[1]Autonomous Systems Lab, Swiss Federal Institute of Technology, Lausanne, Switzerland
[2]Signal Processing Institute, Swiss Federal Institute of Technology, Lausanne, Switzerland

Abstract

The main task of a voice-enabled service robot is to engage people (users) in dialogue providing an efficient access to the services it is designed for within a reasonable (limited) time. In managing service dialogues, extracting the user goal (intention) for requesting a particular service at each dialogue state is the key issue. In service robots deployment conditions speech recognition limitations with noisy audio input and uncooperative users may jeopardize user goal identification. In order to reduce the risk of communication failures due to incorrect user goals, we introduce sequential dialogue repair techniques motivated by the theory of grounding in conversation, and exploiting the inherent multimodality in the perceptual system of a service robot. The error handling method is based on Bayesian networks fusing audio and non-audio based modalities during user goal identification and serving as input to graphical models known as decision networks. Decision networks allow the definition of dialogue repair sequences as actions, and provide an explicit strategy for selecting actions. The paper demonstrates how the above graphical models can be used for designing and implementing of repair strategies for avoiding communication failures in spoken dialogues with mobile tour-guide robots in mass exhibition conditions. The benefits of the proposed repair strategies are tested through experiments with the dialogue system of RoboX, a tour-guide robot successfully deployed at the Swiss National Exhibition (Expo.02).

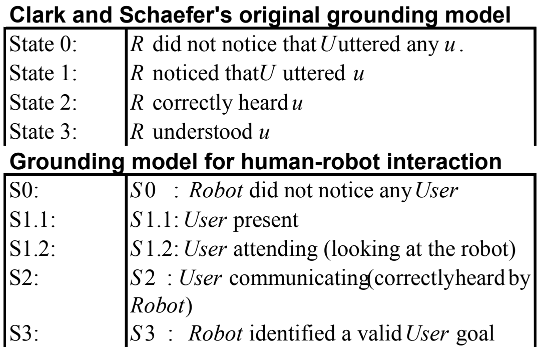
1. Introduction

Mobile service robots are physical agents that act in the physical world, using their mobility to perform tasks useful for humans. These robots perform some fixed number of services. These services can be, e.g. exhibit presentations in the case of mobile tour-guide robot or object delivery in the case of robot assistants. Depending on the service tasks, the robots are equipped with manipulators and sensors forming their multimodal perceptual system. To decide which service to perform service robots need to communicate with their users. Intuitiveness and usability are among the main criteria when designing a communication interface for the perceptual systems of mobile service robots. Speech is an intuitive communication means for humans and therefore service robots employ automatic speech synthesis and recognition in performing their communication tasks. Speech-based interaction requires speech recognition as one of the main input modalities in the robot's system, while different range finders and video cameras are required for safe navigation. The main issue in the spoken service-task oriented dialogue with users is to infer, using speech recognition, which particular service the user intends to request at each dialogue state. The service robot dialogue can then be constructed from a fixed set of states, where the number of possible services defines the state space. Each dialogue state usually contains a spoken interaction in which the robot needs to infer the user goal (e.g. its intention of requesting a given service) in order to decide what service to perform.

The robot's mobility is a generic task for mobile service robots, adding challenges as well as advantages in the spoken interaction. Most of the service robot applications take place in open spaces, where speaking people other than the user and the robot equipment itself corrupt the audio space with high levels of noise. The speech in the input audio signal can originate from users or from other people speaking (passers by) causing errors in speech recognition. Additionally, the end users can be ordinary people lacking any prior robotic experience. They can decide to leave the robot at any time, since this type of interaction is typically short-term. We argue that in such conditions it is preferable that the service robot take the initiative in the spoken dialogue [4]. As stated above people interacting with service robots do not always act cooperatively during the dialogue [21]. Moreover, there are reported cases when they even try to confuse the robot for fun, e.g. misbehaving visitors in a tour-guiding scenario [4], [21]. Such behaviours make visitors' intentions difficult to anticipate in robot-guided human-robot interaction, causing ambiguity and errors when the robot has to interpret them. Communication failures may arise in dialogue due to the above outlined factors. Hence, a service robot managing spoken dialogue with people should employ special care in handling the acquired audio input in order to reduce the effect of unreliable speech recognition, resulting from the adverse acoustic conditions or the audio input contributed from uncooperative users.

In [15], we have outlined the advantages of statistical modality fusion for correcting speech recognition errors in the process of identifying user goals in a tour-guide dialogue scenario. At each dialogue state a Bayesian network was used to elicit probabilities over the possible user goals including an undefined user goal signalling possibility for a dialogue failure. Given the probability distribution over the possible user goals and in particular the undefined one, the error handling strategy should be used to decide explicitly whether to consider the current audio signal unreliable and apply a dialogue repair sequence. If dialogue repair sequences are defined as actions that the service robot can perform, principles from decision theory provide an explicit way for selecting repair actions, given the robot's preferences and the level of uncertainty in the user goal identification at each dialogue state. Based on decision theory, we can define action selection strategies relying on explicit measures of the robot's preferences for actions, i.e. utilities and the principle of maximum expected utility (MEU) [14], [18]. Graphical models related to Bayesian networks known as decision networks are used for implementing utility-based decisions.

In this paper, we report on the use of Bayesian and decision networks for modelling multimodal repair strategies fitted to the requirements of speech-based interaction with a service robot. The paper is structured as follows: Section 2 reviews work in the field of dialogue repair based on grounding in conversation. Section 3 provides background information on decision networks. Section 4 and 5 describe graphical models fusing different modality information to be used for grounding in the context of tour guiding. Section 6 presents how the MEU principle can be used to resolve communication failures resulting from processing audio signal that may be insufficiently grounded (understood) in the tour-guide dialogue. Finally, Section 7 and 8 concludes the paper.



Clark and Schaefer's original grounding model

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

Grounding model for human-robot interaction

S0:	$S0$: *Robot* did not notice any *User*
S1.1:	$S1.1$: *User* present
S1.2:	$S1.2$: *User* attending (looking at the robot)
S2:	$S2$: *User* communicating (correctly heard by *Robot*)
S3:	$S3$: *Robot* identified a valid *User* goal

Figure 1: *State model of grounding in conversation*

2. Dialogue repair and grounding theory in conversation

Spoken human-robot interaction is not only a smooth process of encoding and decoding well-formed acoustical messages. Misunderstandings occur even in the case of correct speech recognition in human dialogues and are collaboratively resolved by the 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. In a model for collaborative contributions to conversation presented in [2] (Figure 1) there are four possible states that an addressee R can attribute to a speaker U and an utterance u. In Figure 1 the states after [2] are also elaborated to cope with human-robot interactions as well. The need for conversational repair (grounding) arises whenever R has failed to reach one of these states given the evidence he has about the other participant. In the case of a speech-based dialogue system the audio signal itself should provide sufficient evidence for inferring the states in Figure 1. All consecutive states have to be reached and failure to reach a given state requires a repair action. Then, the methods for error handling in the domain of human-computer dialogue can be seen as acts of grounding [19] and attributed to states in the model. Traditionally, the dialogue error handling methods use recognition scores to detect recognition errors, correcting the resulting errors through repair dialogues [19], [12], [14]. However, in the case of mobile service robots, detecting errors using only speech recognition can be difficult and repair dialogues may be inefficient in the typical acoustic conditions of robot deployment. Alternatively, a "collaborative" service robot can benefit from the inherent robot multimodality. For example, a "Search for user" repair sequence can be attributed to $S0$ in the Figure 1.

The evidence for failure or success to reach a given state is provided by the robot's input modalities. For example failure to reach state $S3$ can be attributed to the "undefined user goal" (UG=0) modelled by a Bayesian network [15] fusing information from the robot sensors. The strength of evidence about this event can be quantitatively estimated by the posterior distribution of the event "undefined user goal" (UG=0), i. e. $P(UG=0|E=e)$, given the evidence $E=e$ from the input modalities. Bayesian networks can model all the states in Figure 1 and repair actions can be then defined for each state. In modelling the actions selection strategy, we use concepts from decision theory, i.e. utilities.

3. Decision theory

In decision theory and artificial intelligence the principle of maximum expected utility (MEU) is used for modelling the strategy of action selection of an intelligent utility-driven agent [18]. Such an agent maintains an internal state (model) of its environment given its sensors' information. A utility function is used to model the agent's preferences for the different actions through which the agent can manipulate its environment. The utility function assigns a numerical value to each agent's action, given the current state of the environment. Finally, the process of action selection is modelled by combining principles from probability and utility theory. Probability theory is used to model the agent's internal state, given the information (evidence) extracted from its sensors. Utility theory is used to model the agent's preferences between the states of the external environment resulting from a decision taken (executed action). These preferences are captured by the utility function as mentioned above. We use utility function $U(s, a)$ to denote the utility of an action a given that the agent is in a state s. $P(S=s|E=e)$ denotes the probability of each state value, given the current evidence $E=e$ from the sensor data. Then the maximum expected utility is defined by the following equation [17], [14]:

$$MEU(A \mid E) = \arg\max_{a} \sum_{s} P(S = s \mid E) \cdot U(s, a) \qquad (1)$$

The maximum expected utility principle in decision theory states that an intelligent agent should choose the action that maximizes the expected utility of that action, given the sensor evidence for the state of the world at the instant of decision-making. This kind of utility-driven decisions can be implemented with the help of decision networks [14]. In a decision network (*DN*) there are three types of nodes, i. e. chance nodes (ovals), decision nodes (rectangles) and utility nodes (diamonds). An example of a decision network is shown in Figure 3. The chance nodes represent random variables. The agent is usually uncertain about the exact values of these variables. Some of the chance nodes can represent features extracted from the agent sensors; others can represent different aspects of the agent internal state. Decision nodes represent possible choice of actions. The utility nodes represent the utility function. Since the utility function depends on the agent's internal state and the actions, utility nodes usually have one or more chance nodes and the decision node as parents. Bayesian networks are often used to model the probabilistic dependences between the chance nodes and serve as an input to the decision network. A Bayesian Network (*BN*) [13], [17] is a graphical model used to describe dependences in a multivariate probability distribution function (pdf) defined over a set of random variables. The topology of the network is defined by a *D*irected *A*cyclic *G*raph (*DAG*), consisting of nodes corresponding to the variables and arcs representing the conditional dependence assumptions between the variables. The arcs point in the direction from the cause to the consequence or from the parent variable to its children. Thus, Bayesian networks specify a family of statistical models, equipped with a unified set of efficient algorithms for inference [13], e. g. computing posterior probability over set of "query variables", given an assignment for some set of observed (evidential) variables in the network. Therefore, Bayesian networks can produce the probability values on the state variables, i. e. $P(S{=}s|E)$ for the utility-driven agent. Then applying (1) will result in selecting the action with MEU, given the set of possible actions. To construct a decision network for a particular decision problem a precise definition of the agent's internal state, actions and preferences (utilities) is derived from the requirements of the agent's task.

4. Service robot dialogue

We assume that service robot dialogue systems offer a limited set of services that are among the possible communicative goals of the user. The dialogue is organized as a sequence of question/answer states in which the service robot takes the initiative. In the rest of this paper we take as an example the service robot RoboX [7]. This robot was designed to provide tour-guiding services and it was successfully deployed at the Swiss National Exhibition (Expo.02) [4], [8]. For the purpose of human-robot interaction, RoboX is equipped with the following modalities: speech recognition system, interactive buttons, and video camera as input modalities, and LED matrix animations, expressive face, speech synthesis system as output modalities. For the purposes of navigation and obstacle avoidance the robot is additionally equipped with 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 [7]. During Expo.02 eleven robots interacted with individual visitors as well as crowds of people. The question/response pair in the case of RoboX is at the beginning of each exhibit's presentation and consists of yes/no question from the robot and answer from visitor. One complete tour during Expo.02 was limited to five question/response pairs (user goal/actions decision points). Given the main task of the tour guide, e.g. presenting exhibit information, the average number of exhibit presentations, resulting from correctly recognized responses, can be used as a measure for successful interaction.

During Expo.02 there were often cases when people did not follow the choice suggested by the robot, using out-of-vocabulary words and even giving both *yes* and *no* answers or providing no answer at all [4]. 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 *O*bserved *R*ecognition *R*esult *ORR*={*yes*, *no*, *GB*} is then mapped into three possible user goals (*UG*), accounting for the visitor intention: "the user is willing to see the next exhibit" (*ORR*=yes then *UG*=1); "the visitor is unwilling to see the next exhibit" (*ORR*=no then *UG*=2) and "user goal is undefined" (*ORR*=*GB* then *UG*=0). During Expo.02 background "babble-like" acoustic noise and uncooperative visitor's behaviour caused significant errors in recognizing the *GB* word. This was the case for example when initially interested visitors were leaving the robot to respond to other people speaking to them. When this behaviour was coinciding with the question/response pair, the *GB* word was often misrecognized for *yes* or *no* answer by the robot. In order to infer the right user goal (*UG*=0) in this case auxiliary information from the laser scanner signal revealing presence of visitors in close distance with respect to the robot's microphone array (<1.5m), proved to be beneficial [4]. It should be noted however that the laser scanner data provide insufficient information as far as presence of a communicating user is concerned. For example, different objects in the environment, such as for example chairs or the back of a user who is not interested in the robot at all can cause particular patterns that can be easily interpreted as legs. In such cases information from the robot's camera can be very useful.

5. Bayesian networks in the grounding model for service robots

The above Expo.02 experiences showed that in the case of cooperative user behaviour during interaction a particular sequential pattern in the robot input modalities was observed First, the laser modality provides a particular "leg pattern", while at the same time the attention of the user is grabbed and detected by a presence of a frontal face in the video modality, and finally speech is detected by the speech modality and the recognition system detects a "valid" user goal (UG={1,2}). This sequence serves as motivation for constructing the grounding state sequence *S*0-*S3* in Figure 1. We associate the binary event *UR*=1 to the state *S*1.1 and *UR*=0 to *S*0, where *UR*=1 means "User in Range” as detected by the laser modality in the underlying Laser Scanner Signal *LSS*). The event *UA*=1 (User Attending as detected by the video and laser modality) to state S1.1. *UC*=1 (User Communicating as detected by speech, video and laser modality) to state *S*2. At the end the binary event “valid User Goal” (*UG*={1,2} is associated to S3. The speech recognition result (*ORR*) can be seen as the “acoustics-related” aspect of the user goal. On the other hand, to define the influence of the acoustic environment on the speech recognition reliability we define the binary event *SDR* “data reliability” (*SDR*=1 acoustic data is reliable meaning that *UG*=*ORR*, *SDR*=0 acoustic data is unreliable). To infer the state of *SDR* the tour guide robot needs additional evidence about changes in the environment that can affect the reliability of the incoming data and in particular the effect of acoustic noise on the speech signal. The likelihood (*Lik*) of the observed recognition result along with an estimate of the speech-to-noise ratio (*SNR*) of the captured acoustic signal can provide information about the environmental acoustic conditions [5].

Figure 2 depicts a causal model for the above events (*UR*, *UA*, *UC*, *UG*, *SDR*). The real state of each of the events is never known with certainty by the robot. Hence, the causal influences depicted in the figure should be seen as probabilistic. Bayesian networks are widely agreed to provide optimal probabilistic representation for quantifying causal influences and providing inference about probabilistically related events. The cause-effect ordering *UR*, *UA*, *UC*, *UG*, accounts for the state ordering in the grounding model for human-robot interaction (Figure 1) as well as the actual time sequential ordering between the events. In order to have any conversation, first the event “user in range” (*UR*=1) has to appear. It precedes the event “user is attending” (*UA*=1) and user is communicating (*UC*=1). Similarly *UC*=1 precedes the event valid or undefined user goal *UG*={2,1,0}. The observed recognition result *ORR* is the effect of *UG*={1,2} or can be a result from environmental conditions causing unreliable speech recognition (*SDR*=0). Since the variables (*UR*, *UA*, *UC*, *UG*, *SDR*) are not observed during the robot operation, we need to provide additional sources of information that can be observed and can provide evidence in favour of a particular variable state. The laser scanner reading *LSS* can provide evidence in favour of *UR*=1, the result of a frontal Face Detection (*FD*) system can be seen as the effect from *UA*=1, while in addition the event *UC*=1 will need evidence from an indicator of “Acoustic Signal Detected” *ASD*. It was already stated that the likelihood (*Lik*) of speech recognition and a speech-to-noise related measure *SNR* can carry information correlated with the event *SDR*={0,1}. We add to its evidential variables the *ASD* indicator as well.

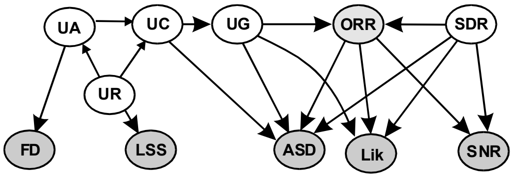


Figure 2: *Final Bayesian network (BN)*

The full set: *E*={*LSS*, *FD*, *ASD*, *Lik*, ***SNR***, *ORR*} of the observed (evidence) variables is shaded in Figure 2. The BN arcs represent the cause/effect relations as outlined above. In order to perform consistent inferences given the set *E*, the parameters of the conditional probability distribution for the network variables have to be learned from training examples. The Bayesian network conditional probability distributions in our case are probability tables for the discrete variables and single conditional Gaussians for the continuous ones. Since video data from Expo.02 were not available, to collect data for all the three modalities we have designed a new dialogue scenario similar to the ones at Expo.02. In this scenario RoboX is playing the role of an interactive Lab tour-guide in the corridor of the Autonomous System Lab (ASL) at EPFL providing presentation of the laboratory activities using posters in an informal conversation.

Multimodal dataset: During each question/answer pair the *E*={*LSS*, *FD*, *ASD*, *Lik*, ***SNR***, *ORR*} values were acquired. The data acquisition sequence is done in 3 phases. First, a recording of 16 bits PCM audio at sampling rate of 16 kHz and RGB video signal at resolution of 320x240 for 2 s (2 s is the assumed duration of the visitor’s answer) is performed. The frame rate of the video signal is approximately 15 fps. After the initial audio-visual acquisition phase a laser scanner reading is acquired from the SICK scanner and the speech recognition is run in the final phase. In that way values for the observed variables *E*={*LSS*, *FD*, *ASD*, *Lik*, ***SNR***, *ORR*} in the Bayesian network are acquired. *LSS* is a continuous value derived after accumulating and normalizing the laser reading samples within a sector of 20^0 with respect to the robot’s front. *FD* and *ASD* are binary variables accounting for the event of detected face during 6 consecutive frames (*FD*=1) and an audio signal detected (*ASD*=1) as reported by a speech activity detector. The face detector in our case is based on the modified algorithm of Viola and Jones [20], while the speech activity detection and subsequent *SNR* calculation was done in a similar way as in [17]. The data were gathered on-line during real communication with people and hidden variables were manually tagged afterwards. To avoid confusion it should be noted that due to implementation issues the data acquisition sequence does not directly correspond to the actual (physical) order of appearing of the events (*UR*, *UA*, *UC*, *UG*, *SDR*) as described in the causal model (Figure 2).

Normal dataset: 33 individual users (21 male, 12 female) were recorded in normal “collaborative” sessions and 18 “non-collaborative” sessions, resulting in 408 question/answer data samples of the form *V*={*UG*, *SDR*, *UC*, *UA*, *UR*, *LSS*, *FD*, *ASD*, *Lik*, ***SNR***, *ORR*}. During the “non-collaborative” sessions we have simulated typical uncooperative user behaviours matching failures to reach the different states in Figure 1. We have recorded “empty” scenarios (failure in S0 Figure 1) in which the robot was circulating in the corridor acquiring data samples without any user at all. Other scenarios corresponded to failures in states S1.1 and S1.2 (Figure 1) such as users that are presenting their backs to the robot during the data acquisition or users that stay and look at the robot without speaking. Finally users were encouraged to use out-of-vocabulary words or speak to others using yes/no answers that might mislead the speech recognizer.

Degraded dataset: In addition to the normal dataset, we acquired degraded data containing babble-type additive noise for 2 male users in 51 different sessions. This dataset has sessions where users are in range or not (*UR*=1 or 0), facing the robot or not (*UA*=1 or 0), speaking or remaining silent (*UC*=1 or 0), and finally using either yes/no answers, or staying silent or using out-of-vocabulary answers (*UG*={1,2} or 0). In all cases, during the sampling of the user's answer the robot's loudspeakers were used to play back babble-like crowd noise recorded during real use of the robot at Expo.02. The average *SNR* for the degraded dataset was computed to be 1.4 dB.

BN training and testing: The BN was trained on 2/3 of the data samples using maximum likelihood technique [13], and tested with the remaining ones. A cross-validation technique similar to the one used in [17] was used to randomly select mutually exclusive parts of the data for training and testing. To test the accuracies of the individual grounding state predictor variables (*UR*, *UA*, *UC*, *SDR* and *UG*) we ran 2 sets of 50 cross-validation tests, where the training and testing portions of the datasets were chosen at random each time. Values for the posteriors *P*(*UR*|*E*), *P*(*UA*|*E*), *P*(*UC*|*E*), *P*(*SDR*|*E*) and *P*(*UG*|*E*) were obtained from the Bayesian network (Figure 2) for each testing sample *E*. Values for the corresponding state predictor variables were assigned using *argmax* criteria on the corresponding posterior probabilities. Tests were done for the events *UR*=1, *UA*=1, *UC*=1, *SDR*=1, *UG*={1,2} (valid user goal) computing corresponding accuracies. The accuracy of *ORR*={yes/no} (baseline speech recognizer) was also calculated and compared with that of *UG*={1,2}. Results can be seen in Table 1.

Predictor	Acc %	std %
UR=1	81.9	4.3
UA=1	85.7	2.5
UC=1	84.4	3.6
SDR=1	82.7	3.5
UG={1,2}	84.3	3.5
ORR={Y/N}	77.2	5.4

(a)

Predictor	Acc %	std %
UR=1	78.1	3
UA=1	85.1	2.8
UC=1	64.4	3.5
SDR=1	68.4	3.2
UG={1,2}	66.6	3.3
ORR={Y/N}	35.7	2.9

(b)

Table 1: *Accuracies for the grounding state predictors with the normal dataset (a) and the combined dataset (b).*

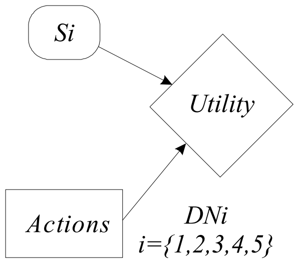
As can be seen in Table 1 (a), the baseline recognizer output (*ORR*) has slightly lower accuracy than the *UG* (user goal) predictor. This can be explained by the fact that the *UG* predictor is better at classifying "garbage" cases (*UG*=0). The second set of tests was performed on the combined dataset consisting of normal and degraded datasets, where training and testing samples could come randomly from any dataset (Table 1 (b)). Still, the *UR*, *UA*, *UC* and *SDR* indicators function above chance level in these noisy conditions and can be put to use in managing a repair sequence. On the combined dataset, the baseline recogniser's performance drops significantly. In this case, the use of *UG* inference will provide more reliable data than straight speech recogniser output.

6. Decision networks for tour-guide repair strategies

In the context of a utility driven tour-guide robot the tour-guide dialogue can then be seen as a process of decision-making, where each state in dialogue is considered as a decision point. At each decision point the "question/response" pair is used to probe the external environment and elicit a probability distribution over the robot's internal states - *P*(*S*|*E*). In the previous section we have equipped the tour-guide robot with models for its internal states, e.g. a Bayesian network for estimating *P*(*S*|*E*), where *S*=*UG*, and *E*={*LSS*, *FD*, *ASD*, *Lik*, ***SNR***, *ORR*}. In order to apply (1), we still need to define precisely the set of robot's actions and the utility function.

Defining actions and repair strategies: The dialogue sequences presenting the exhibits in one complete tour can be seen as valid dialogue actions for the case of *UG*=1. We will refer to these sequences as "Present exhibit" actions. On the other hand, the question/response pairs offering exhibit presentations to the visitors can be seen as valid actions for the case of *UG*=2. We will refer to these actions as "Offer another exhibit" actions. Due to uncooperative visitors and the adverse acoustic conditions during dialogue the visitor's intentions cannot always be classified into meaningful user goals in the context of tour guiding (e.g. simple accept/reject responses in the case of RoboX). In this case, using an "undefined" user goal (*UG*=0) is well motivated and requires "repair" actions for avoiding communication failures. To define the "repair" actions, we take into account the tour-guiding dialogue requirements, i.e.: provide exhibit information through efficient speech-based interaction in limited time, where the number of presented exhibits, after correct user goal identification, can be used as a measure for efficient interaction. Dialogue repair sequences generally occur as unexpected sequence in the normal process of human robot interaction and may lead to delays in communication. Given the tour-guide task requirements the "repair" actions should avoid unnecessary repetitive patterns that might arise using speech recognition in noisy acoustic conditions. Therefore in building "time-saving" repair sequences using alternative input and output robot modalities can be very beneficial. For example, in the case of an absence of communicating visitor (*UR*=0, *S*0 in Figure 1) the most appropriate repair sequence should involve the robot mobility to search for a visitor. We define such a repair sequence as the "Search for visitor" action. In the case of (*UR*=1, *UA*=0 S1.1 in Figure 1) we perform an "Attract visitor" repair, e.g. moving the camera and/or rotating the robot in order to attract the attention and detect the face of the user. Whenever (*UR*=1, *UA*=1, *UC*=0 *S*1.2. in Figure 1) we define a "Hint User" action in which the user who is estimated to be attending but not contributing any audio input is provided with a hint, e.g. the possible ways of answering to the robot. Finally, if the user is communicating (*UC*=1) a "Repeat repair" action, e.g. asking the user for repeated input trial would be the fastest possible repair sequence. However knowing that *UC*=1 and *SDR*=0 would give less motivation to the use of speech-based "Ask for repeat" repair action, compared with an alternative use of the interactive buttons through the "Offer buttons" repair action. In real conditions however the states of *UG*, *UC*, *UA*, *UR* and *SDR* are never known with full certainty. Hence, *UG*, *UC*, *UA*, *UR* and *SDR* are seen as chance nodes and decision networks can be used as a state transition models for selecting valid actions using the principle of maximum expected utility (MEU - Equation (1)).

Decision networks for tour-guide repair strategies: Figure 3 depicts the decision networks *DN*1, *DN*2, …,*DN5* that can be used for selecting actions in the five decision



Actions (a)
1 Present next exhibit
2 Offer another exhibit
3 Repair

S1=UG

Utility *U(a,s)*		*User goal (UG) state* (s)		
		0	1	2
Action (*a*)	1	1	3	2
	2	1	1	4
	3	4	1	2

Actions
1 Next state repair
2 Search for visitors

S2=UR

Utility *U(a,s)*		*UR state*(s)	
		0	1
Action (*a*)	1	1	2
	2	3	1

Actions
1 Next state repair
2 Attract visitors

S3=UA

Utility *U(a,s)*		*UA state*(s)	
		0	1
Action (*a*)	1	1	3
	2	2	1

Actions
1 Next state repair
2 Hint user

S4=UC

Utility *U(a,s)*		*UC state*(s)	
		0	1
Action (*a*)	1	1	2
	2	2	1

Actions
1 Ask for repeat
2 Offer buttons

S5=SDR

Utility *U(a,s)*		*SDR state*(s)	
		0	1
Action (*a*)	1	2	5
	2	5	3

Figure 3: Decision networks (*DN*1, *DN*2, *DN*3, *DN*4, *DN*5) for the tour-guide dialogue transition diagram.

levels of the tour-guide dialogue in Figure 1. The same Bayesian network (Figure 2) is used as an input for the three decision networks to output values for the corresponding posterior distributions needed for equation (1), e.g. $P(S|E)=P(UG|E\ )$ in the main dialogue sequence case *DN*1, $P(S|E)=P(UR|E)$ for the 1st level; *DN*2, $P(S|E)=P(UA|E)$ for the 2nd level *DN3*, etc. and finally *DN*5, $P(S|E)=P(SDR|E)$ for the 4th level of dialogue repair, given the evidence $E=\{LSS, FD, ASD, Lik, \mathbf{SNR}, ORR\}$. The utility functions associated with the utility nodes in the five decision networks that are used in calculating the corresponding actions' expected utilities are defined as real valued tables, indexed by actions and user goals. The numerical values of utilities in general are mathematically unique up to a positive affine transformation [14]. The particular $U(s, a)$ values in the utility tables for the above decision networks represent the tour-guide preferences over its actions, given the user goal values and are motivated by the tour-guiding requirements. For example, due to the time limit during interaction the most preferable action for a "rational" tour-guide robot would be to "Present next exhibit" in the case of *UG*=1, and the least preferable one would be the "Repair" action, since it might lead to unjustifiable delays in interaction. However, in the case of *UG*=0 performing the "Repair" action would be much more relevant in order to prevent communication failure. Given the utility tables, formula (1) can be used by the five decision networks in the order specified in Figure 3 to select the actions that maximize the expected utility of that action, given the distribution over the values of the corresponding chance nodes $\{UG, UC, UA, UR, SDR\}$.

Testing: The combined dataset has been used to train the *BN* in Figure 2. Overall 544 examples were used for training and 272 unseen examples were provided to the *BN* in Figure 2 for inference with roughly half of the cases (154) being undefined user goal (*UG*=0). In order to test the benefits of the proposed repair strategies, we have performed tests with only the data for *UG*=0. We have used the posteriors $P(UG|E)$, $P(UR|E)$, $P(UA|E)$, $P(UC|E)$ and $P(SDR|E)$ calculated by the BN in Figure 2 for 154 cases of an undefined user goal (*UG*=0). The decision network *DN*1 was used initially to decide if a repair action is needed. In that case the rest of the decision networks (*DN*2, *DN*3, *DN*4 and *DN*5) were used to decide if there is a visitor in front of the robot, if the visitor is attending and communicating (speaking to the robot) and consequently if the audio input was reliable (correctly heard). Repair actions were taken according to Figure 3. The results from the experiment are shown in Table 2.

7. Discussion

As can be seen from Table 2 in 96% of the cases the network *DN*1 has correctly assigned a repair action, and 65% of the repair actions correspond to "Search for visitors" actions. At the end there are 3 cases in which the user was estimated to be present and he/she is reoffered to use the speech modality as a repair action in the final repair level. Given that visitors might utter out-of vocabulary words at that point the "Ask for repeat" action may lead to delays in conversation. To handle this issue making the utilities dependent on the number of times an action is executed (e.g. $U_t < U_{t\text{-}1}$) might be beneficial [14]. Finally, in 148 out of 154 cases the mobility of the tour-guide robot provides an efficient way to avoid sure communication failure due to the absence of visitor during interaction.

Main sequence (DN1)		**MEU action**	**corr %**
Actions:	1 Present next exhibit	6	4.6%
	2 Offer new exhibit	0	0.0%
	3 Repair	148	96.1%
1st level repair (DN2)		**MEU action**	
Actions:	1 Next state	96	64.9%
	2 Search for visitors	52	35.1%
3rd level repair (DN3)		**MEU action**	
Actions:	1 Next state	62	65%
	2 Attract visitors	34	35%
4th level repair (DN4)		**MEU action**	
Actions:	1 Next state	49	79%
	2 Hint User	13	21%
5th level repair (DN5)		**MEU action**	
Actions:	1Repromt User	3	6%
	2 Offer buttons	46	94%

Table 2: *Experimental results: Main sequence correctness, 1st …5th level repair action percentage*

Decision theoretic repair strategies provide substantial degrees of freedom in modelling the tour-guide behaviour. Given equally likely user goals the MEU principle will select the action with the maximal sum of the utilities across all user goals (the sum of the rows in the utility tables). In that sense the individual $U(a, s)$ values also contribute to the global importance (preference) on actions. Following such global preferences the behaviour of the tour-guide robot during interaction can be adapted to be more or less conservative in performing the repair actions. For example the global importance of presenting exhibits can be adjusted to be higher compared with the one of offering a new exhibit or the repair option. Since searching for visitors might encourage the visitors around the robot to join the interaction, the global preference is in the favour of the "Search for visitors" action in the first level of the tour-guide repair strategy. In the decision network corresponding to the last repair level (Figure 3), i.e. "Ask for repeat" vs. "Offer buttons" the second action can be seen as globally more preferable. Since buttons input during speech-based interaction is not affected by the acoustic noise, it is considered as more reliable at high levels of acoustic noise.

8. Conclusion

In this paper, we presented a methodological concept for designing and implementing repair strategies for avoiding communication failures in spoken dialogues with mobile service robots. The repair strategies were motivated by general principles from the theory of grounding in conversation and fitted to the requirements of a particular service robot task, a tour-guide robot in mass exhibition conditions. In these conditions the non-collaborative visitors' behaviour and the adverse acoustic conditions were shown to be among the main factors for communication failures in speech-based interaction. The problem of tour-guide dialogue management was shown to depend on a robust inference of the user goal at each dialogue state, where the chance for communication failure can be explicitly modelled through an "undefined user goal". Bayesian and decision networks were used to elicit dialogue repair sequences in accordance with the tour-guide requirements, exploiting the potential benefit of different input and output robot modalities. Decision network implementing the MEU principle allowed us to model complex task-oriented tour-guide behaviours through manipulation of the utility function values. It was shown that decision networks could be used for modelling a variety of tour-guide repair strategies, taking into account different aspects of the user goal.

References

[1] Brennan, S. E. and Hulteen, E. A. 1995. "Interaction and feedback in a spoken language system: a theoretical framework." Knowledge-Based Systems 8: 143-151.

[2] Clark, H. H., and Schaefer, E. F. (1989). "Contributing to discourse." Cognitive Science, 13:259-294

[3] Burgard, W. et al., 1999. Experiences with an interactive museum tour-guide robot, *Artificial Intelligence*, **114** (1-2), pp. 1-53.

[4] Drygajlo, A.et al., 2003. On developing voice enabled interface for interactive tour-guide robots, *Advanced Robotics*, **17** (7), pp. 599-616.

[5] Huang, X., Acero, Al., Hon, Hsiao-W., 2001. *Spoken Language Processing*. Prentice Hall PTR.

[6] Horvitz, Er., Paek, T., 1999, A computational architecture for conversation. *Proc. of the 7th Int. Conf. on User Modeling*, Banff, Canada, June 1999, pp. 201-210.

[7] Jensen, B. et al., 2002a. The interactive autonomous mobile system RoboX. *Int. Conf. on Intelligent Robots and Systems*, *IROS* 2002, Lausanne, Switzerland, Sept. – Oct., 2002, pp. 1221-1227.

[8] Jensen, B., et al., 2002b. Visitor flow management using human-robot interaction at Expo.02. *Workshop: Robotics in Exhibitions, IROS* 2002, Lausanne, Switzerland, Oct. 2002.

[9] Jensen, F., 1996. *An Introduction to Bayesian Networks*. UCL Press.

[10] Jensen, F., Lafferty, J. D., Mercer, R. L., 1990. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly*, **4**, pp. 269-289.

[11] Keizer S. et al., 2002. Dialogue act recognition with Bayesian networks for Dutch dialogues*. Proc. of 3rd SIGdial Workshop on Discourse and Dialogue*, Philadelphia, PA, 2002.

[12] McTear, M., F.,2002. "Spoken Dialogue Technology: Enabling the Conversational User Interface". ACM Computing Surveys **34** (2002) 90–169.

[13] Murphy, K., 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. thesis, U. C. Berkeley, July 2002.

[14] T. Paek, Er. Horvitz, "On the utility of decision-theoretic hidden subdialogues," *Proc of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, Chateu-d'Oex-Vaud, Switzerland August 28-31, 2003.

[15] Prodanov, Pl., Drygajlo, A. 2005. "Bayesian Networks Based Multimodality Fusion for Error Handling in Human-Robot Dialogues Under Noisy Conditions," to appear in ISCA journal of Speech Communication, 2005.

[16] Prodanov, Pl., et al., 2002. Voice enabled interface for interactive tour-guide robots. *Int. Conf. on Intelligent Robots and Systems, IROS* 2002, Lausanne, Switzerland, Sept. – Oct., 2002, pp. 1332-1337.

[17] Richiardi, J., Prodanov, Pl., Drygajlo, A., 2005 "A probabilistic measure of modality reliability in speaker verification," Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICCASP 05, Philadelphia, USA, March 18-23, 2005.

[18] Russell, St., Norvig, P., 2003. *Artificial Intelligence A Modern Approach*. Prentice Hall.

[19] Skantze, G., 2003. "Exploring human error handling strategies: Implications for spoken dialogue systems". ITR Workshop on Error Handling in Spoken Dialogue Systems, Chateau d'Oex, Vaud, Switzerland, August 28-31, 2003, pp. 71-76.

[20] Viola P., Jones M., 2001."Rapid Object Detection Using a Boosted Cascade of Simple Features", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), ISSN: 1063-6919, Vol. 1, pp. 511-518, December 2001.

[21] Willeke, T., Kunz, C., Nourbakhsh, I., 2001. The history of the Mobot museum robot series: An evolutionary study. *FLAIRS* 2001, May, 2001.