Bayesian Networks for Error Handling through Multimodality Fusion in Spoken Dialogues with Mobile Robots

Plamen Prodanov[1], *Andrzej Drygajlo*[2]

[1]Autonomous Systems Lab, [2]Signal Processing Institute,
Swiss Federal Institute of Technology, Lausanne, Switzerland
{Plamen.Prodanov, Andrzej.Drygajlo}@epfl.ch

Abstract

In this paper, we introduce Bayesian networks architecture for combining speech-based information with that from another modality for error handling in human-robot dialogue system. In particular, we report on experiments interpreting speech and laser scanner signals in the dialogue management system of the autonomous tour-guide robot RoboX, successfully deployed at the Swiss National Exhibition (Expo.02). A correct interpretation of the user's (visitor's) goal or intention at each dialogue state under the uncertainty intrinsic to speech recognition accuracy is a key issue for successful voice-enabled communication between tour-guide robots and visitors. Bayesian networks are used to infer the goal of the user in presence of recognition errors, fusing speech recognition results along with information about the acoustic conditions and data from a laser scanner, which is independent of acoustic noise. Experiments with real-world data, collected during the operation of RoboX at Expo.02 demonstrate the effectiveness of the approach in adverse environment. The proposed architecture makes it possible to model error handling processes in spoken dialogue systems, which include complex combination of different multimodal information sources in cases where such information is available.

1. Introduction

Mobile robots can move freely in their operational environment and are often used to perform tasks through close interaction with humans. One typical example is the mobile tour-guide robot, whose task is to engage the visitors in a tour, guiding them, moving autonomously, and presenting the items of the exhibition (exhibits). Recent efforts were reported for introducing voice-enabled interfaces as the most intuitive form of interaction with visitors to mass exhibitions [2], [13]. The tour-guide robot equipped with speech synthesis and recognition systems allows spoken interaction. One interaction cycle, which consists of a given number of exhibit presentations can be determined in advance from the particular exhibition plan [6], [13]. The tour-guide dialogue is then constructed from a set of states, where the number of possible exhibit presentations per tour defines the state space. At each state the task of the dialogue system is to infer the goal of the visitor through the speech recognition component, i.e. his/her intention to attend the possible next state presentations in order to decide on which exhibit to present next.

The operating conditions in a mass exhibition environment abound with a variety of uncertainties [1], [19]. Visitors' intentions are difficult to anticipate in human-robot interaction [2], causing ambiguity and errors when the robot has to interpret them. Data coming from the robot's input modalities, in particular the speech signal captured by the microphone can be very noisy. The presence of a crowd of people and moving robots in the exhibition room results in adverse acoustic conditions, causing errors in the speech recognition. Hence, a system managing speech-based interaction with visitors should employ error-handling techniques in order to avoid communication failures.

Standard techniques for error handling in speech recognition are based on detecting errors using the recognition scores and correcting them through repair dialogues [3]. Detecting errors using only speech recognition can be difficult and repair dialogues may be inefficient in the acoustic conditions of mass exhibition. The type of interaction faced by a tour-guide robot in the exhibition room is usually short-term, as visitors want to see as many exhibits as possible. They do not have prior knowledge in robotics and typically the initiative in the dialogue is left to the robot then. In such conditions people hardly tolerate repetitive, time-consuming repairs that may often occur in the noisy conditions and will most probably drive the visitors away. The shortest delay in communication can be associated by an alternative method that detects and corrects errors immediately. In this paper we report on one such method, based on Bayesian networks and error handling through multi-modal signal fusion, using auxiliary information from acoustics-insensitive signals to compensate for speech recognition errors.

The paper is structured as follows. Section 2 describes the error handling in human-robot dialogue based on multimodality fusion in inferring the intention of the visitor (user goal). In Section 3 Bayesian networks are introduced as a probabilistic framework for fusing the speech and laser modalities in inferring the goal of the visitor. In Section 4 the approach is tested through experiments with real data, collected during the deployment of the tour-guide robot RoboX at the Swiss National Exhibition Expo.02 [5]. Finally the potential benefits of multimodality fusion for error handling in spoken dialogues with robots are outlined with future perspectives in the Discussion and Conclusion parts of the paper (Sections 5 and 6).

2. Error handling in human-robot dialogue

Speech recognition errors are common in human-robot speech communication. Even simple utterances (single words) pronounced in adverse (noisy) acoustic environment are frequently misclassified or just missed in the recognition process. It is often said that the major problem in voice-enabled human-robot interfaces is their inability to detect and correctly handle different speech recognition errors. Thus, error management in real-world spoken dialogue applications is crucial for successful human-robot interaction. However,

most of the current tools for speech application development do not have decent support for error management and their performance depends only on speech recognition error rate in given conditions [18].

Error handling in its three main phases, e.g. error prevention, detection and correction [18] can take place in all human-robot interaction levels. Some recognition errors can be prevented through controlling the robot initiative context dependant grammar. Such methods lead to less natural but more robust speech recognition performance. Detection of errors can be done using special models for out-of-vocabulary words, e.g "garbage models" or can be based on the recognition score - accounting for the confidence of the recognizer in its output. Finally the error-correction component is usually a small dialogue, which can include a yes/no confirmation question or explicit request for repetition. In this case, in exhibition noisy conditions, we can end up in repetitive error correction combinations. Signalling misunderstanding through these frequent error corrections can be very frustrating and can give the user an impression of a dialogue failure [17]. To prevent such failures, when dealing with the robot's speech input we can reduce errors by analyzing the ambient acoustical environment and using information from other input modalities (laser scanner, video camera or touch buttons) through multi-modal signal fusion.

Multi-modal signal fusion is generally defined as any method that combines different signals to perform inferences that may not be possible from a single signal [16]. If the inference aims at finding the intention of the speaking visitor based on unperfect speech recognition, multi-modal signal fusion can be seen as an efficient method for error handling in human-robot dialogues. In order to define precisely the meaning of the visitor's intention and how inference can be performed through multi-modal signal fusion, some specific details concerning the dialogue system of a tour-guide robot are needed.

2.1. Human-robot dialogue

We take as an example the interactive tour-guide robot RoboX successfully deployed at the Swiss National Exhibition Expo.02. During Expo.02 the tour-guide robot interacted with individual visitors as well as crowds of people (hundreds of thousands of visitors during 5 months, seven days a week, 10 hours per day). They were not instructed beforehand as how to operate the robot. In such conditions it is preferable that the tour-guide robot takes the initiative in the spoken dialogue [2]. Thus, a successful tour-guide robot should be capable of detecting the presence of people, engaging them in dialogue, presenting the items of the exhibition (exhibits). During this dialogue the visitors' intentions and behaviour can vary from collaborative to investigative and even "destructive" [2]. The tour-guide robot needs to interpret this behaviour into "user goals" relevant to the tour-guide dialogue. The tour-guide dialogue can be represented as a set of dialogue states, where each dialogue state corresponds to a sequence of low-level behavioural events, such as a speech synthesis event, a speech recognition event, a robot movement event, etc. The sequence of events forming a dialogue state is organized to present a specific exhibit. Thus the number of dialogue states is fixed and can be defined in advance based on the number of exhibits described by the particular exhibition plan. Each dialogue state contains verbal interaction in the form of initiative/response pair, during which the speech recognition is typically used to infer the "goal" of the speaker in the context of the current state [4]. We assume that the spoken utterances coming from visitors during interaction can be mapped into a finite number of state dependent user goals, which are used to infer the next dialogue state. In Figure 1 this process is depicted graphically; *UG* stands for the user goal and *DS* for the dialogue state. We assume that the state of the dialogue at time *t* depends on the dialogue state and the user goal at time *t*-1, and it can also affect the current user goal at time *t*. Then the key issue in spoken dialogue management is to decide on the most likely user goal into the current dialogue state.

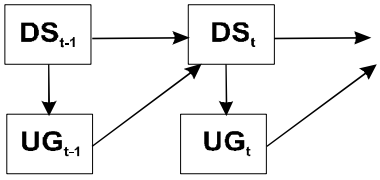

Figure 1: *Dependency graph for spoken interaction management*

The initiative/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; e.g. the tour-guide robot asks the visitors if they want to see the next exhibit. One complete tour consists of five presentations [6]. A successful dialogue 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 typical visitors' behaviour. Since the visitors' behaviour during the dialogue can vary as reported in [2] there are often cases when people do not follow the choice suggested by the robot, using out of vocabulary words and even giving both *yes* and *no* answers or simply remain silent. Therefore the speech recognition system of RoboX needs 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 mapped then 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 the recognition system was additionally trained with noisy speech recorded from the real visitor answers. This resulted in improved recognition performance. However the background exhibition noise caused significant errors in recognizing the *GB* word as can be seen from the row *ORR* Acc (observed recognition result accuracy) in Table 1 (Section 4). This was the case for example when initially interested visitors were leaving the robot to respond to other people calling them. When this unexpected behaviour was coinciding with the initiative/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 (<1.5m), facing the microphone array is potentially beneficial.

2.2. Multi-modality fusion for error handling

The speech recognition result can be seen as the "acoustics-related" aspect of the user goal at the given dialogue state. The laser scanner signal provides information about the location of the communicating visitor with respect to the microphone of the robot. This "spatial" aspect of the user goal is essential, as absence of a user in given range in front of the robot could signal possible communication failure. To incorporate the information about presence of people for spoken communication explicitly we define the binary event *U* "user in range for communication" (*U*=1 user is in range, *U*=0 user is out of range). Combining the observed recognition result (*ORR*) with evidence from the noise independent *L*aser *S*canner *Signal* (*LSS*) that can affect the event *U*=1 can change the "confidence" about the result of speech recognition. To define the influence of the acoustic environment on the speech recognition reliability we define the binary event *DR* "data reliability" (*DR*=1 acoustic data is reliable, *DR*=0 acoustic data is unreliable). To infer the state of *DR* 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 [3]. The *U* and *DR* events can directly influence *ORR*. Since this influence is not directly established the causal relationships should be seen as probabilistic. It can happen that people are near the robot, speaking as expected and recognition errors still occur.

Bayesian networks have been shown to perform inference about probabilistic events compatible with the notion of causal reasoning [7]. They have recently emerged as a promising tool for fusing multiple sources of information in dialogue modelling and pattern classification [4], [9], [10], [11].

In the sections that follow, the concept of Bayesian networks is presented. Bayesian networks are used to fuse data coming from the speech recognition and the laser scanner modalities of RoboX for inferring the user goals of the visitors.

3. Bayesian networks

Bayesian Networks (BNs) are graphical models used to describe a joint probability distribution (pdf) over a finite set of random variables [12]. The pdf structure is defined 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. Independence assertions provide a compact representation for the joint pdf through factorising it into product of independent conditional terms. The arcs between the nodes point from parent variables to their children variables. The intuition behind directionality represents a cause-effect relationship. From the probabilistic point of view the arcs converging to 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 joint pdf in the general case of *N* variables $V = (X_1, X_2, \ldots, X_N)$ can be derived from the chain rule for the probabilities, after declaring the conditional independence assumptions given by the network's topology:

$$P(X_1, X_2, ..., X_N) = \prod_{i=1}^{N} P(X_i \mid X_{i-1}, ..., X_1)) = \prod_{i=1}^{N} P(X_i \mid Parents(X_i)) \quad (1)$$

Parents(X_i) are all the parent nodes for X_i. The equalities: $P(X_i \mid X_{i-1}, ..., X_1) = P(X_i \mid Parents(X_i))$ declare the conditional independence assumptions encoded in the BN's graph. *Parents*(X_i) specifies the set of possible causes of X_i or the set of variables that can directly influence the variable X_i.

The nodes' variables in a Bayesian network can be discrete or continuous. Probability tables represent the CPD for a discrete variable. Arbitrary parametric CPDs can be assigned to the continuous ones. Conditional Gaussian distributions appear as a frequent choice in modelling continuous variables [10]. One reason for this option is that if all arcs are allowed except those pointing from a continuous parent to a discrete child the resultant BN is a multivariate conditional Gaussian [15]. Gaussian mixture models (GMMs) appear as a special case of such a distribution given that the continuous variables have only discrete parents [10]. At the same time the Bayesian network encodes efficiently and intuitively the parameter space of the model through the dependence assumptions and their cause/effect interpretation [15].

3.1. BN properties

Three basic connections can exist in a general Bayesian network, e.g. serial, diverging and converging (Figure 2). Evidence provided by an instantiated variable can pass through a serial or diverging connection, until the intermediate variable is not instantiated. In the case of converging connection the evidence in one of the parent nodes can affect the state of the other only if their common child is instantiated. These properties are referred as the rules of "d-separation", which can be used to infer local conditional independences among the variables [7]. If evidence can pass from some variable to another taking certain path in the network, the two variables become dependent.

In the example for a convergent connection in Figure 2 the set of network's variables *V* = (*U*, *DR*, *ORR*) consists of three discrete variables. Their conditional probability distributions (CPDs) are simply tables containing the discrete values for the probabilities *P*(*U*), *P*(*DR*) and P(*ORR*|*U*, *DR*). The joint pdf in this case can be written as: *P*(*U*, *DR*, *ORR*) = *P*(*U*)*P*(*DR*)*P*(*ORR*|*U*, *DR*). The event of presence of visitor is associated with the variable *U* and the event that the observed speech recognition result (*ORR*) can be unreliable is associated with the variable *DR* (*DR*=0). These two events can be seen as direct causes that can influence the particular value of *ORR*. Indeed if *yes* or *no* keyword was recognized our belief about presence of a visitor communicating with the robot rises, and certain suspicion that speech recognition can be incorrect appears. Now we acquire additional evidence coming from the laser scanner signal that there is really a user in close distance with respect to the microphone. In most of the cases this is done for the purpose of communicating. Thus this new evidence reduces our initial belief that speech recognition result is unreliable (erroneous). The presence of user has

explained away the observed recognition result and has lowered our suspicion on the performance of speech recognition in noisy conditions. Such way of inter-causal relationship, commonly known as "the explaining away phenomenon" [7] can be numerically encoded in the BN's CPDs and demonstrated using inference.

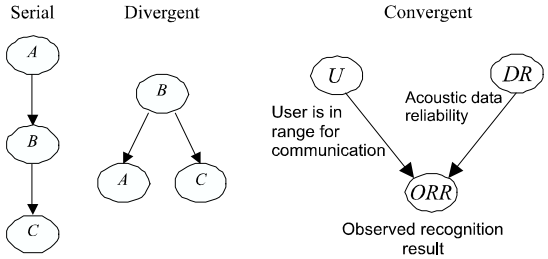

Figure 2: *Basic Bayesian network structures*

3.2. Inference in Bayesian networks

The basic task for any probabilistic inference system is to compute the posterior distribution for a set of "query" variables, given some observed event, i.e. an evidence for some observed variables. In the inference problem we calculate $P(X_K|Y)$, where $X_K \subseteq X$ is the subset of query variables from the full set of unobserved variables $X = \{x_0,\ldots,x_{L-1}\}$; $Y = \{y_0,\ldots,y_{M-1}\}$ is the subset of the observed (evidential) variables and $V_N = X \cup Y = \{v_0,\ldots,v_{N-1}\}$ is the set of all N variables. Once the CPD functions for all the nodes given their parents are defined, an exact or approximate inference on each node in the network can be done [10], [12]. In the simplest and least efficient case exact inference can be performed through marginalizing the full joint pdf after entering the evidence value. Finally, in order to perform consistent inference, estimates for the conditional probability distribution parameters have to be learned from training examples for the network variables.

3.3. CPD parameter learning

The goal of the CPD parameter learning is to obtain estimates for the conditional distribution functions of the variables from data (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 initialisation and a maximum likelihood (ML) training technique. During the training the CPD parameters are adjusted in order to maximize the likelihood of the model with respect to the training data examples (Appendix C.2 in [10]).

3.4. Decision making using inference

The inferred posterior distribution $P(X_K|Y)$ for the query variable X_K can be used for making decisions on a particular value for X_K, based on the observed evidence $Y=e$. If X_K is a discrete variable this last step can be seen as a classification problem in which X_K is the classification variable. Different optimality criteria for assigning X_K to one of its possible class values exist. To select the most likely X_K we use an argmax criterion.

$$X_k = \arg\max_{x_K} (P(X_K = x_k \mid Y = e)) \quad (2)$$

4. BNs for *UG* inference

To build a Bayesian network model for inferring the most likely user goal (*UG*) value, we need first to define precisely the set of random variables, the conditional independence assumptions between them and the variables of interest for inference: $P(X_K|Y)$. In our case $X_K = UG$. We define the network variables' set as V = (*UG*, *DR*, *U*, *ORR*, *LSS*, *Lik*, *SNR*). *UG* and *ORR* are ternary, while *U* and *DR* are binary discrete variables - accounting for the user goal, observed recognition result, the presence of user for communication and the acoustic data reliability as described in Section 2. *LSS*, *Lik* and *SNR* are continuous variables. The *Lik* value is obtained from the speech recognizer of the robot. *SNR* is an estimate of the speech-to-noise ratio of the current speech signal in dB. *LSS* is a two-dimensional vector, calculated from the signal generated by the laser scanner. The details on computing the continuous variables' values are given in section 4.1.

In defining the set of arcs A between the set of variables we follow the cause/effect intuition. In the diagrams for convenience purposes, discrete variables are drawn as squares, and ovals are used for the continuous ones. To mark an observed variable we will use shading.

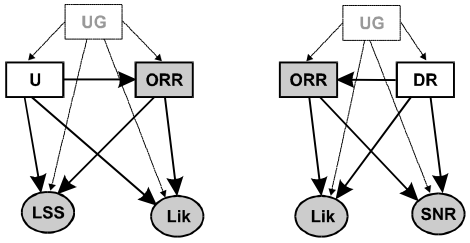

Figure 3: *Bayesian networks for* a) *user presence and* b) *data reliability based UG inference*

The Bayesian networks depicted in Figure 3 encode: a) the causal influence of the presence of a user in front of the robot (*U*); and b) the causal influence of the acoustic data reliability (*DR*) on the observed recognition result *ORR* and the corresponding data variables associated with them - *LSS*, *Lik* and *SNR*. The implicit assumption behind the Bayesian network in Figure 3 a) is that whenever there is a user near the microphone (*U*=1) he is most probably speaking. User's presence influences the *LSS* value and user's speech influences the recognition result (*ORR*) and its likelihood (*Lik*). On the other hand acoustic data reliability (DR) can be a cause for errors in speech recognition and for low *Lik* and *SNR* values. Finally the user goal *UG* is seen as the direct cause for the values of all the other variables, so we add the corresponding arcs (dashed lines in Figure 3).

The example for converging connection that was presented in Section 3.1, Figure 2 can give the intuition behind the inter-causal relationship between the two main causes (*U*, *DR*) for the observed recognition result *ORR*. Hence the Bayesian network for fusing the speech and laser scanner modality can be obtained through unifying the two networks from Figure 3. Figure 4 depicts the final form of the Bayesian network built with the set of variables V for the purpose of the user goal classification. Shaded variables are observed during the

inference of $P(X_K|Y) = P(UG|LSS, Lik, ORR, SNR)$, where $X_K=UG$ and $Y= (LSS, Lik, ORR, SNR)$.

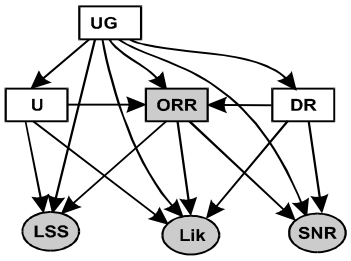

Figure 4: *Final Bayesian network for user goal classification*

4.1. Training of the BNs

The training examples are taken from real data (audio files and laser scanner readings), collected during the deployment period of RoboX at Expo.02. The audio files contain a speech signal, sampled at 16 kHz, with duration of 2 seconds, corresponding to the duration of a yes/no answer. The laser scanner signal contains a sequence of values corresponding to distances to obstacles in the environment (walls, humans, etc.) reflecting the laser beam of the scanner. Within an angle interval of 360^0 and 0.5^0 precision, the laser signal contains 722 distances in meters (m) with resolution of 0.5 mm [5]. Only the values within the interval $[255:285]^0$ are taken in order to account for presence of visitors in range for spoken interaction (the event *U*). This angle sector corresponds to the front of the robot, where the microphone array is located. To eliminate noisy reflections and to reduce the dimensionality of the resulting vector, we divide this interval into two equal parts, adding the distances contained in them, and normalizing the resulting values by the length of the intervals. The resulting two-dimensional vector is used as the variable *LSS* in the Bayesian network. *ORR* and *Lik* values are obtained from the recognizer of the robot. According to its definition, *DR*=0 when *ORR* does not match with *UG* and *DR*=1 when *ORR* matches the goal of the user *UG*. As already stated *U*=0 corresponds to the event "there is no user in range for spoken communication" and *U*=1 corresponds to the opposite event. Hence, when *UG*={1,2} then *U*=1. Finally, values for the *SNR* are estimated from the speech files [14]. We assume that single Gaussians can model the values of the continuous variables. For training the BN models (Figure 3, 4), we use 270 training examples for each value of *UG*, resulting in 810 sequences of the form: {*UG, U, LSS, DR, Lik, ORR, SNR*}.

4.2. Testing

For testing the model, we use 130 testing examples per given value of *UG*, resulting in 390 testing sequences. After training the network, we perform Bayesian inference on *UG*, given the evidence from the samples of testing data on *LSS*, *Lik*, *SNR* and *ORR*. Since the Bayesian network has only 7 variables, we use a method of exact inference based on the junction tree algorithm [8]. Using this algorithm, a value for $P(UG|Y) = P(UG=ug \mid ORR=o, Lik=l, SNR=sn, LSS=[d_1,d_2])$ is calculated for each $ug \in \{0,1,2\}$ and every testing sample $s=\{o, l, sn, [d_1,d_2]\}$. The result from the experiment for the final BN (Figure 4) is depicted graphically in Figure 5. The first curve shows the real values for *UG* from the testing samples and the other three curves show the values for $P(UG|Y)$ inferred by the network. To select the most likely user goal we use the criterion (2), where $X_K=UG$. Results for the percentage of accurately classified cases, using the BNs in Figure 3 and 4 (BN in Fig. 3 (a), BN in Fig. 3 (b), Final BN Acc), compared to the accuracy of the observed recognition result (ORR Acc) are given in Table 1.

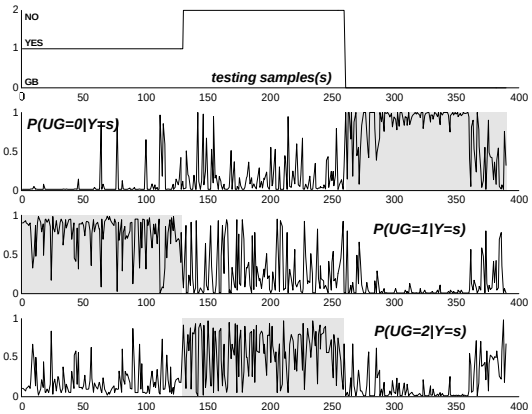

Figure 5: *Graphical representation for P*(*UG*|*Y*).

Table 1: *UG identification: ORR and BN accuracy*

UG	0	1	2	Overal
ORR Acc	38.5%	93.1%	66.9%	66.2%
BN in Fig. 3 (b)	77.7%	84.6%	61.5%	74.6%
BN in Fig. 3 (a)	78.4%	86.9%	66.9%	77.4%
Final BN Acc	**80.8%**	**85.3%**	**67.7%**	**77.9%**
Gain	42.3%	-7.8%	0.8%	11.8%

5. Discussion

The comparison in Table 1 shows a significant improvement in the accuracy of the user goal identification, when introducing information from the laser-related aspect and the acoustics reliability aspect using a Bayesian network classifier. The system is avoiding speech recognition errors without any dedicated repair dialogue technique. The gain in performance is due to the improved identification of the garbage case *UG*=0, which in turn is due to the dependences found between the laser scanner data and the speech recognition result in the Bayesian network architecture presented in Figure 4. The rules of "d-separation" can be used to justify these dependences theoretically. The observed testing results confirm this phenomenon as well. For example, in the region corresponding to the undefined user goal *UG*=0 (the shaded region in the second top most plot in Figure 5) the Bayesian network has calculated the following probabilities:

$P(UG=0|s_1) = 0.94$, $P(U=1|s_1) = 0.06$, $P(DR=0|s_1) = 0.06$, for the testing sample: s_1 = (*ORR*='*GB*', *LSS*=[4.8 4.6]m, *Lik*=-71.3, *SNR*=7.8dB); and

$P(UG=0|s_2) = 0.90$, $P(U=1|s_2) = 0.09$, $P(DR=0|s_2) = 0.94$, for the testing sample: s_2 = (*ORR*='yes', *LSS*=[4.8 4.1]m, *Lik*=-67.2, *SNR*=1.2dB).

In both the examples people are far from the tour-guide robot (> 4m). In the first case the recognizer has correctly spotted a garbage word - *GB*, while in the second case there is an incorrectly recognized *yes* word. Although the higher likelihood in the second case, the low probability of user presence - $P(U=1|s)$, and the low *SNR* value (giving rise to the probability of unreliable acoustic data - $P(DR=0|s_2)$ provide

evidence in favour of the right decision about the most likely user goal - *UG*=0. In the case when *UG*={1,2} there is not any significant gain in using the Bayesian network, which is an intuitive result as the laser scanner does not provide information for distinguishing between the words *yes* and *no*. Additionally when people are close to the robot and the models for speech recognition were trained with noisy speech in average conditions the results for *yes* and *no* can be even slightly degraded (*UG*=1 or 2). Possible benefit in this direction can result from using information from a video camera images tracking the lip-movement of the communicating speaker.

The results presented in the third and forth row of Table 1 outline the relative importance of the additional information extracted from the "*U"* (user presence) related and "*DR"* (data reliability) related data. It can be seen that introducing information from the laser scanner signal leads to greater benefits, compared with the case when only auxiliary information concerning the acoustic data reliability is used.

Finally, to evaluate users' general satisfaction from the improved *UG* identification, we have implemented a scenario in which RoboX is presenting posters to visitors of the Autonomous Systems Lab, EPFL. A version using the BNs described above will be implemented as well. User's satisfaction will be then analyzed trough questionnaires using criteria similar to these of the visitor survey in [2].

6. Conclusion

In this paper we introduced a new approach for error handling in spoken dialogue systems for mobile tour-guide robots working in mass exhibition conditions. The problem of dialogue management was shown to depend on a robust inference of the user goal at each dialogue state. While the process of identifying the user goal only from the speech recognition result can be inefficient in noisy exhibition conditions, using the additional acoustic noise-insensitive laser scanner signal can be beneficial. The framework of Bayesian networks was introduced for detecting and correcting errors in the user goal classification problem using multimodal input. We have demonstrated that a Bayesian network can model efficiently the dependencies between the speech and the laser scanner signals. In addition, the method allows the explicit modelling of the speech recognition reliability enabling the possibility to exploit both the strengths and the weaknesses of the speech recognizer in deciding for the true user goal. The performance of the model was tested in experiments with real data from the database, collected during the deployment period of the tour-guide robot RoboX at Expo.02. The results show that the Bayesian networks provide a promising probabilistic framework for error handling in multimodal dialogue systems of autonomous tour-guide robots.

7. References

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

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

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] Nefian, Ar. Et al., K., 2002. Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, **11** (1-15).

[12] Pavlovic Vl. I., 1999. *Dynamic Bayesian Networks for Information Fusion with Application to Human-Computer Interfaces*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

[13] 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.

[14] Prodanov, Pl., Drygajlo, A., 2003. Bayesian networks for spoken dialogue management in multimodal systems of tour-guide robots. *Proceedings of the* 8th *European Conference on Speech Communication and Technology, Eurospeech* 2003, Geneva, Switzerland, September 2003, pp. 1057-1060.

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

[16] Smith, M., 2003. *Bayesian Sensor Fusion: A framework for Using Multi-modal Sensors to Estimate Target Locations & Identities in a Battlefield Scene*. Ph.D. thesis, Department of Statistics, Florida State University.

[17] 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.

[18] Turunen, M., Hakulinen, J., 2001. Agent-based error handling in spoken dialogue systems. In *Proc. Eurospeech* 2001, pp. 2189-2192.

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