Multimodal Interaction Management for Tour-Guide Robots Using Bayesian Networks

Plamen Prodanov
Autonomous System Lab
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
Plamen.Prodanov@epfl.ch

Andrzej Drygajlo
Signal Processing Institute
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
Andrzej.Drygajlo@epfl.ch

Abstract – In this paper, we propose a Bayesian network framework for managing interactivity between a tour-guide robot and visitors in mass exhibition conditions, through robust interpretation of multi-modal signals. We report on methods and experiments interpreting speech and laser scanner signals in the spoken dialogue management system of the autonomous tour-guide robot RoboX, successfully deployed at the Swiss National Exhibition (Expo.02). A correct interpretation of a user's (visitor's) goal or intention at each dialogue state is a key issue for successful speech-based interaction in voice-enabled communication between robots and visitors. We introduce a Bayesian network approach for combining noisy speech recognition results with noise-independent data from a laser scanner, in order to infer the visitors' goal under the uncertainty intrinsic to these two modalities. We demonstrate the effectiveness of the approach by simulation based on real observations during experiments with the tour-guide robot RoboX at Expo.02.

I. INTRODUCTION

Recent advances in speech processing and mobile robotics has made it possible to deploy autonomous tour-guide robots with voice interfaces, enabling spoken dialogue with visitors to mass exhibitions [2], [15]. The operating conditions in a mass exhibition environment are abound with a variety of uncertainties [1], [18]. Visitors' intentions are difficult to anticipate in human-robot interaction [16], [19], causing ambiguity and errors when the robot interprets 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 many people and moving robots in the exhibition room results in adverse acoustic conditions, causing errors in the speech recognition [15]. Hence a dialogue system that relies on the output of speech recognition only, in managing human-robot interaction, can result in communication failures due to recognition errors. In such conditions auxiliary information from noise-independent input modalities can be useful [13], [12], [11]. The typical sensing devices used by autonomous robots, such as laser scanners [6] produce data that is unaffected by the acoustic conditions. Visitors' intentions and goals can result in dependent data patterns in the underlying laser scanner reading as well as the speech signal. Under this assumption a multi-modal data interpretation for inferring visitors' goals becomes an attractive prospect. This interpretation can be based on statistical pattern recognition techniques using Bayesian networks. Bayesian networks are known in the domain of artificial intelligence for modeling statistical dependencies between different causes and consequences [8], [12]. They have recently emerged as a promising tool for fusing multiple sources of information in pattern recognition and classification [9], [4], [11], [10].

In this paper we report on multimodal methods using speech and laser scanner signals in spoken dialogue management system of the tour-guide robot RoboX [2], [6], [17]. We use the framework of Bayesian networks to develop robust multimodal interpretation for managing the tour-guide interaction with visitors in mass exhibition conditions.

II. INTERACTION MANAGEMENT FOR TOUR-GUIDE DIALOGUE

The interaction between the visitors and the tour-guide robot in mass exhibition conditions is generally short term, as people wish to see as many exhibits as possible in limited time and are initially unprepared for communication with the tour-guide robot [2], [19]. In such conditions it is preferable that the tour-guide robot takes the initiative in the spoken dialogue interaction [15]. 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 behavior can vary from collaborative to investigative and even "destructive" as reported in [19], [2]. The tour-guide robot needs to interpret this behavior into "user goals" relevant to the tour-guide dialogue. The tour-guide robot dialogue can be represented as a set of dialogue states, where each dialogue state corresponds to a sequence of low-level behavioral events [6], [2], 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.

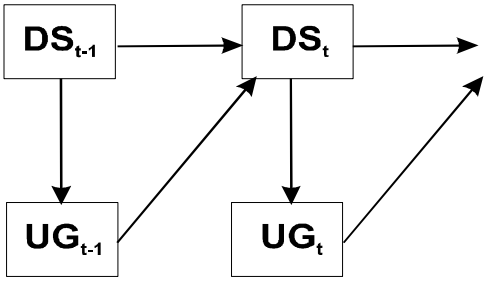

Fig. 1 Dependency graph for spoken interaction management

We assume that the number of dialogue states is fixed and can be defined in advance based on 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 dialogue state [5], [16]. 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 Fig. 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 context 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 interaction management is to decide on the most likely user goal into the current dialogue state.

The role of interaction management during the tour is to infer the visitor's goal in the current dialogue state, using mainly speech recognition, and to supply related exhibition information, according to this goal in the next dialogue state. However, the background exhibition noise can cause speech recognition errors. While combining the Observed Recognition Result (*ORR*) with the noise independent Laser Scanner (*LS*) reading can be beneficial, it is also important for the tour guide robot to sense changes in the environment that can affect the interaction and estimate the reliability of the incoming data. The likelihood (*Lik*) of the observed recognition result along with estimate for the signal-to-noise ratio (*SNR*) of the speech signal in the current dialogue state can give information about the environmental acoustic conditions.

Finally, we need to find the most likely user goal (*UG*) from a fixed number of goals $\{ug_1,\ldots,ug_N\}$ at a given state in the dialogue, having the underlying sequence of (*ORR, Lik, LS, SNR*) data values. The user goals for the interaction between visitors and the tour-guide robot in mass exhibition are defined in section III.A. The probabilistic model, from which the most likely user goal can be determined, can be created using Bayesian networks.

In the sections that follow, the concept of Bayesian networks is presented as an efficient framework for handling the user-goal classification problem in presence of multimodal data coming from the speech recognizer and the laser scanner on the RoboX platform. Experimental results with simulated data based on real-life observations during experiments with tour-guide robot RoboX at Expo.02 are presented in section IV.

III. BAYESIAN NETWORKS

A Bayesian Network (*BN*) is a graphical model used to describe dependencies in a multivariate probability distribution function (pdf) defined over a set of variables. The topology of the network is defined by a Directed Acyclic Graph - DAG. The graph consists 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. Fig. 1 is one example of a Bayesian network. In this network the user goal *UG* at time *t* has one parent variable - the dialogue state *DC* at time *t*. If we define the conditional probability distribution functions for all nodes given their parents, an exact or approximate inference on each node in the network can be done [10], [12]. In the inference problem we want to calculate $P(X_K|Y)$, where $X_K \subseteq X$ is a subset of interest from the set X. $X = \{x_0,\ldots,x_{N-1}\}$ and $Y = \{y_0,\ldots,y_{M-1}\}$, $M+N=L$ denote the two subsets of hidden and observable variables in the set $Z_L = X \cup Y = \{z_0,\ldots,z_{L-1}\}$ of all L random variables in the Bayesian network. $X_K = UG$ in the case of user goal classification.

To build a *BN* model for the user goal we need to define the set of random variables, the conditional dependence assumptions between them and a way to estimate their conditional probability distribution from data.

A. Experimental framework

For the experiments we use the RoboX platform. The set of dialogue states on RoboX platform consists of fixed number of initiative/response pairs, related to the exhibits that the robot presents [7]. The initiative/response pair 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 guide asks the visitors if they want to see the next exhibit. The speech recognizer can 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 *ORR*={*yes, no, GB*} is mapped into three possible user goals: 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). One complete tour consists of five presentations [2]. Successful interaction can be then measured by the average number of correctly recognized responses at the beginning of each exhibit presentation. Therefore the user

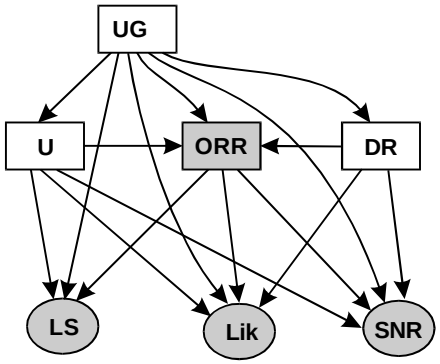

Fig. 2 Bayesian network for user goal classification

goal classification task consists of choosing one of the three user goals given the speech recognition result. This classification can be incorrect in presence of recognition errors. This is often the case with noisy speech recognition in mass exhibition conditions. If such poor recognition performance persists the visitors usually leave the robot [2].

B. Definitions of the variables

To prevent the tour-guide from talking to itself when people have left, we first assume that presence of people for spoken interaction is governed by the hidden variable U (U=1 user is in range, U=0 user is out of range). The range for spoken interaction depends on the microphone array used to capture the speech signal and usually varies between distances of 0.5 and 1.5 m within angle of $20 - 30^0$ [2] with respect to the microphone. After the visitor's response we can observe the recognition result *ORR* and its likelihood *Lik*. We assume that the performance of the recognizer is governed by a hidden variable accounting for speech Data Reliability *DR*. The current values for *ORR*, *Lik* and *SNR* depend on *DR*, i.e. DR=1 corresponds to reliable speech data corresponding to low levels of background noise and an accurate recognition result, while DR=0 corresponds to unreliable speech data that is likely to produce error in the recognition output. *LS* is the laser scanner reading which is a two-dimensional vector corresponding to the most probable location of a visitor in polar coordinates (distance in m and angle in ^{0}) with respect to the robot. Finally, the user goal of the visitor is governed by the *UG* variable, which is hidden and related to all the other variables. These relations represent the conditional dependence assumptions given by the network's DAG.

C. Bayesian network topology

Bayesian networks are usually handcrafted according to the cause/consequence dependence among the random variables. Fig. 2 depicts a Bayesian network built with the set of variables defined in section B for the purpose of user goal classification. The arcs define the causal relations that can be derived from the description of the variables. Square nodes correspond to discrete random variables and round ones to continuous random variables. Shaded variables are observed during the inference. We are interested in inferring $P(X_K|Y)=P(UG|LS, Lik, ORR, SNR)$. We assume that *UG* is the primary cause of all variables. The observed recognition result *ORR* can be additionally affected by *DR* and *U*, and is the direct cause for the observed likelihood of the recognition result *Lik*. *U* can affect the observed laser scanner reading *LS*, which we assume to be also a consequence of the observed recognition result *ORR*. Finally *DR* is the cause for the observed *SNR* that can also be correlated with the *ORR*.

D. Training of the Bayesian network

After defining the network topology we need to specify parameters of the conditional pdfs for the continuous variables. First, we assume that all continuous variables are modeled sufficiently by single Gaussians. Then, to perform inference, we need to estimate the conditional distribution functions of the variables (the conditional probability tables for the discrete variables and the parameters of the Gaussian pdfs for the continuous ones) from data. In the case of full observability of the variables in the training set, the estimation can be done with random initialization and Maximum Likelihood (*ML*) training technique. During the training the pdfs are adjusted in order to maximize the likelihood of the model with respect to the training data examples [10]. In order to supply enough training and testing data examples for the experiment, we perform a simulation. The goal is to model the relation between the hidden and observed variables in the Bayesian network and to evaluate how well these relations can be captured by this network.

IV. EXPERIMENT WITH SIMULATED DATA

A. Description of the simulator

The experimental observations during the deployment of the tour-guide RoboX at the Swiss National Exhibition Expo.02 [2], [17] guided us in creating the simulator.

First, when visitors answer the questions of the robot they are typically within distance from 0 to 2m and angle sector of $20 - 30^0$ with respect to the microphone on the robot. This corresponds to the case when UG={1,2} and U=1. In the case when visitors are not "collaborative" or their goal is undefined (UG=0, U ={0,1}) the laser scanner reading can take all possible distances and angles within the scanner range (0:8m and $0{:}360^0$).

Second, the microphone captures speech signal simultaneously with background exhibition noise. The noise intensity can vary depending on the particular situation. At the beginning of the day, when there are no many visitors the noise level is not significant, while during the peak hours the level of noise increases. We then assume that the final signal captured by the microphone is given by the sum:

$$NS = S + k \cdot N\,, \qquad (1)$$

where S is the clean signal, k is the mixing coefficient determining the level of noise, and N is the background noise signal. When visitors are speaking to the robot the S is normally bigger than N. Then, the mixing coefficient k varies uniformly between 0 and 1.5 for the case of UG={1,2} and between 1.5 and 2.5 in the case of UG=0. In the case of RoboX the duration of the captured speech signal NS is 2 seconds according to the average duration of yes/no answer [2]. Then, in the simulation, clean speech S of the keywords (yes, no) with duration of 2 seconds is mixed with a 2 second of exhibition background noise N. We define the signal-to-noise ratio as:

$$SNR = 10 \cdot \log_{10}(E(S)/E(N)) \qquad \text{[dB]} \ (2)$$

where E(.) stands for the energy. In the acoustic space of exhibition rooms the effect of reverberation causes the sound to decay in approximately exponential fashion [3]. Thus, we assume that the amplitude of the acoustic signal during propagation decreases exponentially with distance, while the effect of the angle can be modeled by a cosine function. The initial amplitude depends on the speech volume, which is visitor dependent characteristic. Then we assume that the acoustic signal propagates from the visitor's location to the location of the microphone according to the following law:

$$S = Si \cdot \cos(q) \cdot K \cdot e^{-cd}\,, \qquad (3)$$

where S is the final amplitude of the signal; Si is the initial amplitude; K is a positive gain coefficient, accounting for the visitor dependent speech volume. To model the effect of directivity of the microphone $K=K1=2$, when $q \in [-90{:}90^0)$ and $K=K2=1$, when $q \in [91{:}270^0)$, where q is the angle in degrees and d is the distance in m; c is a constant that characterizes the fading rate of the signal.

The simulation variables are defined as follows. LS corresponds to laser scanner reading a vector with two continuous components: $LS(1)$ is distance in m and $LS(2)$ is angle in ^{0}. ORR is the observed recognition result when presenting NS to the recognizer. According to section III.B, we define data reliability variable DR=0, when ORR does not match with UG and DR=1, when $ORR=UG$. Similarly U=0 corresponds to the event "there is no user willing to communicate" and U=1 corresponds to the opposite event. If UG=0 then U=0, when we have undefined user goal we assume that the user is not willing to interact with the robot. Otherwise when $UG \in$ {1,2} then U=1 meaning that the user is willing to interact. Finally, the simulation is done in the following order:

1. Fix a value for UG
2. Determine the value of U
3. Generate value for LS uniformly distributed in {[0:2] [-10:10]} if U=1 and in {[1:8] [0:360]} if U=0.
4. Calculate S from (3) and NS from (1).
5. Supply NS to the speech recognizer.
6. Observe Lik and ORR.
7. Determine DR.
8. Calculate SNR from (2).

We are thus able to produce any number of sequences of the form: {UG, U, LS, DR, Lik, ORR, SNR}.

B. Experimental results

The simulator was used to get training data examples. 1000 values for each UG={0,1,2} were generated, resulting in database of 3000 records.

Experiment 1. In the first experiment we use the network from Fig. 2 assuming that the continuous variables have single Gaussian distribution. For training the model we use the first 700 examples from the simulated data for each value of UG, resulting in 2100 training examples. For testing the model, we use the remaining part of the simulated data, resulting in 900 testing examples. These numbers were motivated by our observation that the variance of the overall classification error is below 0.5% within an interval of 100, when the number of the training examples is above 600. Some statistics including the average of $LS(1)$ in m (Dist), $LS(2)$ in ^{0} (Ang), SNR in dB, and observed recognition likelihood (Lik) for the test data are given in TABLE I.

Experiment 2. For the second experiment we decided to use mixture of Gaussians for the continuous pdfs. In order to keep the time of computation reasonable we chose to use 3 Gaussian mixtures for all the continuous variables. In this experiment, the same data as in Experiment 1 is used for training and testing. In this case, we perform additionally EM (Expectation Maximization) training to get estimates for the mixture weights.

TABLE I TEST DATA STATISTICS

UG	Data	Total
0.00	Count of ORR	300
	Average of Dist	3.460745196
	Average of Ang	180.48
GB	Average of SNR	0.895016876
	Average of Lik	-70.87478067
1.00	Count of ORR	300
	Average of Dist	1.030230118
	Average of Ang	5.016666667
YES	Average of SNR	6.675573123
	Average of Lik	-70.107082
2.00	Count of ORR	300
	Average of Dist	1.017436033
	Average of Ang	4.883333333
NO	Average of SNR	5.255437713
	Average of Lik	-69.195607

TABLE II EXPERIMENTAL RESULTS

Experiment 1			
UG	**0**	**1**	**2**
Facc	0.0033	0.1133	0.0200
Acc	0.9933	0.9867	0.8833
ORR Acc	76%	45%	73%
Experiment 2			
UG	**0**	**1**	**2**
Facc	0.0033	0.0633	0.0400
Acc	0.9967	0.9600	0.9367

After training the network, we perform Bayesian inference on *UG*, given the evidence from the samples of testing data on *LS*, *Lik*, *SNR* and *ORR*. Since our Bayesian network has only 7 variables, we use a method of exact inference based on the junction tree algorithm [12]. Using this algorithm a value for $P(UG|Y) = P(UG=ug|ORR=o, Lik=l, SNR=sn, LS=[d,q])$ is calculated for each $ug \in \{0,1,2\}$ and every testing sample $s=\{o, l, sn, [d,q]\}$. The resulting values for Experiment 1 are depicted in Figure 3. 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:

$$ug = \arg\max_{ug}(P(UG = ug \mid Y = s)) \qquad (4)$$

Results for the proportion of accurately and falsely classified cases (Acc / Facc), using the criterion in (4), compared to the accuracy of the *ORR* for the two experiments are given in TABLE II.

V. DISCUSSION

A direct comparison between the accuracies of the speech recognizer and the Bayesian network *UG* classifier (TABLE II) shows significant improvement of the second one (Bayesian network), in both Experiments 1 and 2. Improving the model by using a mixture of three Gaussians for the continuous variables in the network further improves the result as seen from TABLE II.

The poor performance of the baseline speech recognizer can be explained by the high level of background noise in the simulated speech signal as seen from the SNR values in TABLE I. In such conditions a spoken interaction management system based on interpreting only the unimodal speech recognition is less reliable than the alternative approach, based on multimodal speech recognition, using additional noise-independent information from the robot platform.

All the above-presented results are based essentially on simulations. In order to reveal the potential of the proposed methodology with real data, we have performed slightly modified experiment with speech recordings and laser scanner readings taken during the operation of RoboX as a tour-guide at Expo.02. The results and conclusions are reported in [14].

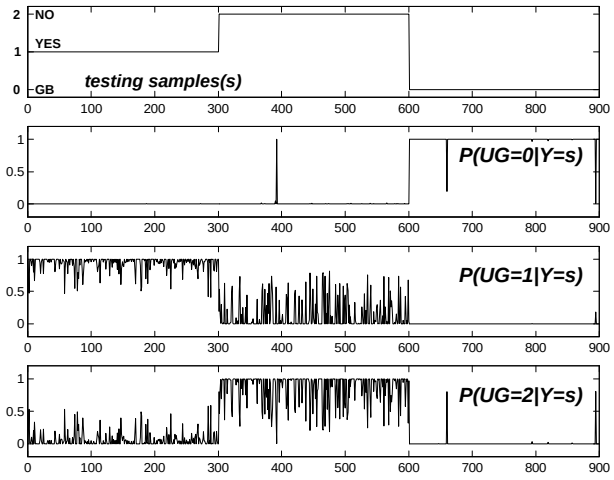

Figure 3 Graphical representation of P(UG|Y)

VI. CONCLUSION

In this paper we introduced an approach for interaction management in mobile tour-guide robots working in mass exhibition conditions. The problem of management of the spoken dialogue interaction was shown to depend on the user goal at each dialogue state. While the process of identifying the user goal from the underlying speech recognition result can be inefficient in noisy exhibition conditions, using the additional noise-independent laser scanner signal can be beneficial. The framework of Bayesian networks was introduced for solving the user goal classification problem using multimodal input. We demonstrated that the dependencies between the speech and the laser scanner signals could be modeled successfully by a Bayesian network. The performance of the model was tested in experiments with simulated data based on real-life observations of the tour-guide robot RoboX during the Expo.02. The results clearly show that the Bayesian networks are promising framework for multimodal interaction management for autonomous tour-guide robots.

VII. ACKNOWLEDGEMENTS

This research work was supported partly by Expo.02 and the Swiss National Science Foundation project “Intelligent Voice Enabled Interfaces for Human Guided Mobile Robots”.

VIII. REFERENCES

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

[2] A. Drygajlo, et al., “On developing a voice-enabled interface for interactive tour-guide robots”, to be published in *Advanced Robotics*, 2003.

[3] B. Gold and N. Morgan, *Speech and Audio Signal Processing*, John Wiley and Sons, 2000.

[4] F. Huang, J. Yang, A.Waibel, "Dialogue management for multimodal user registration," *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP’2000)*, Beijing, China, October, 2000.

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

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

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

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

[9] S. Keizer, R. op den Akker, and A. Hijholt, "Dialogue act recognition with Bayesian networks for Dutch dialogues", *Proc. of 3rd SIGdial Workshop on Discourse and Dialogue*, Philadelhia, PA, 2002.

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

[11] Ara V. Nefian, L. Liang, X. Pi, X. Liu, K. Murphy, “Dynamic Bayesian networks for audio-visual speech recognition”, *EURASIP Journal on Applied Signal Processing*, 11 (1-15), 2002.

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

[13] D. Perzanowski, A. Schultz, W. Adams, and E. Marsh, (2000) "Using a natural language and gesture interface for unmanned vehicles," in *Unmanned Ground Vehicle Technology* II, G.R. Gerhart, R.W. Gunderson, C.M. Shoemaker, eds., *Proceedings of the Society of Photo-Optical Instrumentation Engineers*, vol. 4024, 2000, pp. 341-347.

[14] Pl. Prodanov and A. Drygajlo, “Bayesian networks for spoken dialogue management in multimodal systems of tour-guide robots”, to be published in *Proc. of the 8th European Conference on Speech Communication and Technology, Eurospeech 2003*, Geneva, Switzerland, September 2003.

[15] Pl. Prodanov et al., “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.

[16] N. Roy, J. Pineau and S. Thrun, "Spoken dialogue management using probabilistic reasoning", *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, 2000.

[17] R. Siegwart et al., “Robox at Expo.02: A large scale installation of personal robots”, *Workshop: Robots as Partners: an Exploration of Social Robots IROS 2002*, Lausanne, Switzerland, September 2002.

[18] S. Thrun et al., “Minerva: A second generation museum tour-guide robot”. *IEEE Int. Conf. on Robotics and Automation (ICRA'99)*, Detroit, Michigan, May 1999.

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