# Bayesian Networks for Spoken Dialogue Management in Multimodal Systems of Tour-Guide Robots

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

[1]Autonomous Systems Lab, [2]Signal Processing Institute,
Swiss Federal Institute of Technology, Lausanne, Switzerland
plamen.prodanov@epfl.ch, andrzej.drygajlo@epfl.ch

## Abstract

In this paper, we propose a method based on Bayesian networks for interpretation of multimodal signals used in the spoken dialogue between a tour-guide robot and visitors in mass exhibition conditions. 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 a user's (visitor's) goal or intention at each dialogue state is a key issue for successful voice-enabled communication between tour-guide robots and visitors. To infer the visitors' goal under the uncertainty intrinsic to these two modalities, we introduce Bayesian networks for combining noisy speech recognition with data from a laser scanner, which is independent of acoustic noise. Experiments with real data, collected during the operation of RoboX at Expo.02 demonstrate the effectiveness of the approach.

## 1. Introduction

Recent advances in speech technology and mobile robotics have attracted researchers to introduce voice-enabled interfaces in mobile robotic applications. Today it is possible to deploy autonomous tour-guide robots with such interfaces, enabling spoken dialogue with visitors at mass exhibitions [2, 12]. The tour-guide robot equipped with speech synthesis and recognition systems allows spoken interaction. To perform tasks related to autonomous navigation its mobile platform is typically equipped with a "range-sensing" device such as a laser scanner [5]. The goal of the tour-guide robot is to engage the visitors in a tour, guiding them, when moving autonomously, and presenting the items of the exhibition (exhibits). One full tour consists of a given number of exhibit presentations, which can be determined in advance from the particular exhibition plan [2, 6]. The tour-guide dialogue is then constructed as a sequence of dialogue states, where the number of possible exhibit presentations per tour defines the state space. The task of the spoken dialogue management is to infer the goal of the visitor in the current dialogue state, i.e. his/her intention to attend to the possible next state presentations in order to decide on which exhibit to present.

The operating conditions in mass exhibition environments abound with a variety of uncertainties [1, 14]. Visitors' intentions are difficult to anticipate in the human-robot communication [12, 15], causing ambiguity and errors when the robot interprets them. 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 [12]. Hence, a dialogue system that relies on speech recognition only in managing a human-robot interaction can result in communication failures due to recognition errors. Under such conditions auxiliary information from other input modalities, insensitive to acoustic noise can be useful [11, 10]. The laser scanner produces data that is unaffected by the acoustic conditions. Assuming that visitors' intentions and goals can result in dependant data patterns in the underlying laser scanner reading and the speech signal, a multi-modal data interpretation for inferring visitors' goals becomes an attractive prospect. This data interpretation can be based on statistical pattern recognition techniques using Bayesian networks. Bayesian networks are commonly used in the field of artificial intelligence for modeling statistical dependencies between different causes and effects [7, 11]. They have recently emerged as a promising tool for fusing multiple sources of information in dialogue modeling and pattern classification [8, 3, 10, 9].

In this paper, we report on multimodal methods using speech and laser scanner signals in the spoken dialogue management system of the tour-guide robot RoboX [2, 5]. We use the framework of Bayesian networks to develop a robust interpretation of multimodal signals for managing the tour-guide spoken dialogue with visitors in mass exhibition conditions.

The paper is structured as follows. Section 2 describes the tour-guide dialogue management, based on visitor's goal classification. In section 3 Bayesian networks are introduced as a probabilistic framework for multimodal data interpretation 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.

## 2. Dialogue management for tour-guide robots

The interaction between the visitors and the tour-guide robot in mass exhibition conditions is generally short-term, since people wish to see as many exhibits as possible in a limited time and are initially unprepared for communication with the tour-guide robot [2, 15]. Under such conditions it is preferable that the tour-guide robot take the initiative in the spoken dialogue [12]. Thus, a successful tour-guide robot should be capable of detecting the presence of people, engaging them in a dialogue and presenting the exhibits. During this dialogue the visitors' intentions and behavior can vary from collaborative to investigative and even "destructive" as reported in [15, 2]. The tour-guide robot needs to interpret this behavior as "user goals" relevant to the particular tour-guide dialogue. This dialogue is represented as a series of dialogue states, where each dialogue state corresponds to a sequence of low-level behavioral events [5, 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. 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 a verbal interaction in the form of an initiative/response pair, during which the speech recognition is used to infer the "goal" of the speaker in the context of the current state [4, 13]. We assume that the spoken utterances coming from visitors during interaction can be mapped into a finite number of application specific user goals (*UG*), which are used to infer the next dialogue state. We assume that the state of the dialogue at time *t* ($DS_t$) depends on the dialogue state and the user goal at time *t*-1 ($UG_{t-1}$), and it can also affect the current user goal at time *t*. Then the key issue in the spoken dialogue management is to decide on the most likely user goal at the current dialogue state.

The set of dialogue states of the tour-guide RoboX consists of a fixed number of initiative/response pairs, related to the exhibits that the robot presents [2, 6]. The initiative/response pair is at the beginning of each exhibit's presentation and consists of a yes/no question from the robot and answer from the visitor; e.g. the guide asks the visitors if they want to see the next exhibit. The speech recognition system can distinguish between the keywords e.g. *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 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). Based on this technique the dialogue manager can infer the visitor's goal in the current dialogue state, supplying related exhibition information, according to this goal in the next state. At Expo.02 one complete tour consisted of five exhibit presentations [2]. Successful interaction can then be measured by the average number of correctly recognized responses at the beginning of each exhibit presentation. However, the background exhibition noise can cause speech recognition errors. While combining the observed recognition result (*ORR*) with the noise independent *L*aser *S*canner *R*eading (*LSR*) can be beneficial, it is important for the tour guide robot to sense changes in the environment that can affect the interaction and to estimate the reliability of the incoming data. The likelihood (*Lik*) of the observed recognition result along with an estimate of 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 three possible goals at a given state in the dialogue, having the underlying sequence of (*ORR*, *Lik*, *LSR*, *SNR*) data values. The probabilistic model for determining the most likely user goal can be created using Bayesian networks.

## 3. Bayesian networks for *UG* classification

A Bayesian Network (*BN*) is a graphical model used to describe dependencies 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*). 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.

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 [9, 11]. 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.

### 3.1. Definitions of the variables

The user 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 poor recognition performance persists the visitors usually leave the robot [2]. To prevent the tour-guide from talking to itself when people have left, we first assume that the presence of people for spoken communication is governed by the hidden variable *U* (*U*=1 user is in range, *U*=0 user is out of range). The range for spoken communication depends on the microphone array used to capture the speech signal and is usually specified by a distance between 0.5 and 1.5 m and an angle sector of 20 – $30^0$ [2] with respect to the microphone. The laser scanner reading (*LSR*) 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, *LSR* results in 722 distances in meters (m) with precision of 0.5 mm with respect to the robot [5]. Only the *LSR* values within the interval $[255:285^0]$ are taken in order to account for presence of visitors in range for spoken interaction. This angle sector corresponds to the front of the robot, where the microphone is located. To eliminate noisy reflections and to reduce the dimensionality of the resulting vector, we divide this interval into two equal parts, integrating the distance values contained in them, and normalizing the resulting values by the length of the intervals. The resulting two-dimensional Laser Scanner vector *LS* is used as variable in the Bayesian network. After the visitor's spoken response we can observe the recognition result *ORR*, its likelihood *Lik* and the current laser scanner vector *LS*. We assume that the performance of the recognizer is governed by a hidden variable accounting for speech *D*ata *R*eliability (*DR*). The current values for *ORR*, *Lik* and *SNR* depend on *DR*, i.e. *DR*=1 corresponds to reliable speech data corresponding to low level of background noise and accurate recognition result, while *DR*=0 corresponds to unreliable speech data that is likely to produce error in the recognition output.

Finally, the user goal of the visitor is governed by the *UG* variable, which is hidden and related to all other variables. These relations represent the conditional dependence assumptions given by the Bayesian network's graph.

### 3.2. Bayesian network graph

Bayesian networks are usually handcrafted according to the cause/consequence dependence among the random variables. Figure 1 a) depicts a Bayesian network built with the set of

variables defined in section 3.1 for the purpose of the 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 the 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 vector *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*.

### 3.3. 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 correctly by single Gaussians. Then, to perform inference, we need to estimate 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 initialization and a 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 [9].

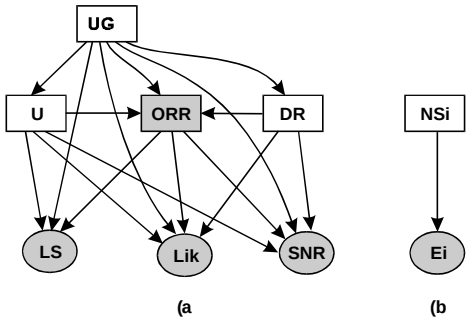


*Figure 1 Bayesian networks for a) user goal classification, b) SNR estimation*

The training examples are taken from real data (speech files and laser scanner readings), collected during the deployment period of RoboX at Expo.02. The files contain a speech signal, sampled at 16 kHz, with duration of 2 seconds, corresponding to the average duration of yes/no answer. *LS* vectors are calculated from the laser scanner readings; *ORR* values are obtained after presenting the speech files to the recognizer of the robot. According to section 3.2, we define the data reliability variable as follows: *DR*=0, when *ORR* does not match with *UG* and *DR*=1 when *ORR* matches *UG*. *U*=0 corresponds to the event "there is no user in range 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 and is not in range for spoken communication. Otherwise when $UG \in \{1,2\}$ then *U*=1, which means that the user is willing to interact. Finally, values for the SNR can be estimated from the speech files.

### 3.4. SNR estimation

In order to estimate the real value for the SNR we need to separate the clean speech and the noise in noisy speech signal in the training data. However this is not trivial, since in the noisy acoustic conditions of the exhibition, the signal from the visitor speaking to the robot has similar characteristics to the background noise, mostly coming from other people speaking. Instead of performing costly calculations to separate speech and noise and calculate the real SNR, we estimate a SNR correlated feature, based on the signal's short-term energy. Short-term energy is calculated using windows containing 400 samples (25 ms) with 50% overlapping. We assume that each energy value in this vector can be generated by two Gaussians distributions, modeling the probability of the current energy value being noise or clean speech segment in the signal. Such a model can be represented in the framework of Bayesian networks as shown in Figure 1 b). $NS_i$ is the hidden variable governing the current energy value being noise or speech, and *Ei* is the current energy value. This network is trained on the speech short-term energy vector, using the expectation maximization algorithm (EM) with random initialization. After training the model, we sample it once again with the energy vector, inferring values for $P(NS_i|E_i)$, where $NS_i=1$ correspond to speech and $NS_i=0$ to noise segments, for each energy component in the vector (Figure 2). At the end the *SNR* correlated feature is defined as follows:

$$SNR = 10 \cdot \log_{10}\left(\frac{\sum_i P(NS_i = 1 \mid E_i)\cdot E_i}{\sum_i P(NS_i = 0 \mid E_i)\cdot E_i}\right) \quad \text{[dB]} \ (1)$$

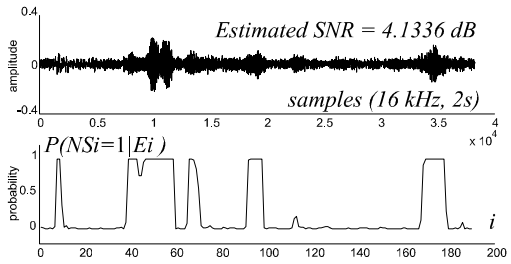


*Figure 2 Experimental results for SNR estimation*

## 4. The *UG* classification experiment

The network from Figure 1 a) is used in the *UG* classification experiment. We assume that the continuous variables have single Gaussian distribution. For training the model, we use 270 training examples for each value of *UG*, resulting in 810 sequences of the form: {*UG, U, LS, DR, Lik, ORR, SNR*}. 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 *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 [11]. Using this algorithm a value for $P(UG|Y) = P(UG=ug \mid ORR=o, Lik=l, SNR=sn, LS=[d1,d2])$ is calculated for each $ug \in \{0,1,2\}$ and every testing sample $s=\{o, l, sn, [d1,d2]\}$. The result from the experiment is depicted graphically 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:

$$ug = \arg\max_{ug}(P(UG = ug \mid Y = s)) \qquad (2)$$

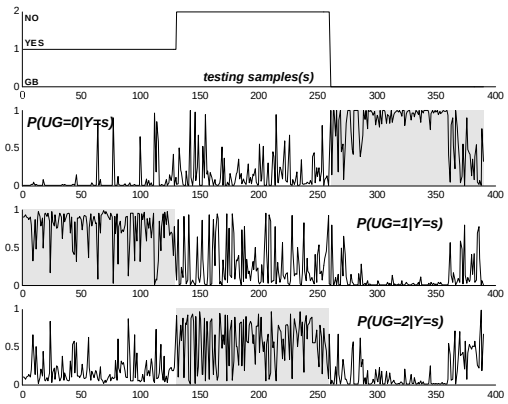


*Figure 3 Graphical representation of P(UG|Y)*

Results for the percentage of accurately classified cases, using the *BN*, and the criterion in (2) (BN Acc), compared to the accuracy of the *ORR* (ORR Acc) are given in Table 1.

| UG | 0 | 1 | 2 | Overall |
|---|---|---|---|---|
| ORR Acc | 38.5% | 93.1% | 66.9% | 66.2% |
| BN Acc | 80.8% | 84.6% | 66.9% | 77.4% |
| Gain | 42.3% | -8.5% | 0.0% | 11.3% |

*Table 1 Experimental results*

The comparison shows significant improvement in the overall performance of the system in the case of introduction of additional laser scanner information and the Bayesian network classifier. The gain is due to the improved classification performance of the garbage case *UG*=0. In the case, 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* answers can be unchanged or even slightly degraded (*UG*=1 and *UG*=2).

## 5. Conclusions

In this paper we introduced a new approach for dialogue management for mobile tour-guide robots working in mass exhibition conditions. The problem of dialogue management was shown to depend on 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 solving the user goal classification problem using multimodal input. We demonstrated that a Bayesian network can model efficiently the dependencies between the speech and the laser scanner signals. 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 result shows that Bayesian networks are a promising framework for dialogue management in multimodal systems of autonomous tour-guide robots.

## 6. 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 Voice Enabled Interface for Interactive Tour-Guide Robots", to be published in Advanced Robotics, Journal of Robotics Society of Japan, 2002.
3. 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.
4. Er. Horvitz, T. Paek, "A Computational Architecture for Conversation", Proc. of the 7th Int. Conf. on User Modeling, Banff, Canada, June 1999, pp. 201-210.
5. 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.
6. B. Jensen et al., "Visitor Flow Management using Human-Robot Interaction at Expo.02", Workshop: Robotics in Exhibitions, IROS 2002, Lausanne, Switzerland, October, 2002.
7. F. Jensen, An Introduction to Bayesian Networks, UCL Press, 1996.
8. S. Keizer, Rieks op den Akker, and Anton Hijholt. "Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues", Proc. of 3rd SIGdial Workshop on Discourse and Dialogue, Philadelhia, PA, 2002.
9. K. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning", Ph.D. thesis, U. C. Berkeley, July 2002.
10. Ar. 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.
11. Vl. I. Pavlovic, "Dynamic Bayesian Networks for Information Fusion with Application to Human-Computer Interfaces," Ph.D. thesis, University of Illinois at Urbana-Champaign, 1999.
12. 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.
13. 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.
14. 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.
15. T. Willeke, C. Kunz, I. Nourbakhsh, "The History of the Mobot Museum Robot Series: An Evolutionary Study", FLAIRS 2001, May, 2001.