Development of Coffee Maker Service Robot using Speech and Face Recognition Systems using POMDP

Widodo Budiharto*, Meiliana** and Alexander Agung Santoso Gunawan***
School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email: * wbudiharto@binus.edu; ** meiliana@binus.edu; *** aagung@binus.edu

ABSTRACT

There are many development of intelligent service robot in order to interact with user naturally. This purpose can be done by embedding speech and face recognition ability on specific tasks to the robot. In this research, we would like to propose Intelligent Coffee Maker Robot which the speech recognition is based on Indonesian language and powered by statistical dialogue systems. This kind of robot can be used in the office, supermarket or restaurant. In our scenario, robot will recognize user's face and then accept commands from the user to do an action, specifically in making a coffee. Based on our previous work, the accuracy for speech recognition is about 86% and face recognition is about 93% in laboratory experiments. The main problem in here is to know the intention of user about how sweetness of the coffee. The intelligent coffee maker robot should conclude the user intention through conversation under unreliable automatic speech in noisy environment. In this paper, this spoken dialog problem is treated as a partially observable Markov decision process (POMDP). We describe how this formulation establish a promising framework by empirical results. The dialog simulations are presented which demonstrate significant quantitative outcome.

Keywords: service robot, coffee maker, speech recognition, face recognition, spoken dialog problem, POMDP

1 INTRODUCTION

In the last few years, service robot technologies can be found in many human environment, both in public and private environment. Therefore, research on service robots has gained growing interest especially in focusing in how to improve the interaction between robot and human being. International Federation of Robots (IFR) defines service robots as a robot which operates semi or fully autonomously to perform services useful to the well-being of humans and equipment, excluding manufacturing operation [1]. Many researches of service robot in different application fields have been conducted such as assistive robot for elderly [2], robotic home assistant [3], and robot waiters in restaurant [2, 4]. This paper will focus on coffee maker service robot as one of promising field that can be applied on restaurant, office or convenient store area.

Based on our previous work [5], the coffee maker service robot will embed speech and face recognition ability to facilitate robot-user interaction. Challenge of similar research area about human-robot interaction is to create natural interaction between user and robot by detecting and recognize speech and face. Face recognition stage dealt with multiple face expressions, accessories wore by user (glasses, etc) and lightness-darkness condition of the images. On the other hand, speech recognition stage considered multi speech source separation and irrelevant noises filtering either for moving or stationary sound sources [6]. Thus the main problem in here is to know the intention of user naturally through spoken interaction. This problem is commonly called as spoken dialog systems (SDS).

SDS allow user to interact with user to know her intention using speech as the primary communication medium [7]. Traditionally, SDS have applied in call centre applications where the aim is to reduce costs by decreasing the requirement for a human operator. Nowadays, the common use of speech interfaces in smartphones demonstrate the value of integrating natural speech interactions into mobile applications. During the last few years, a new approach to dialogue managementhas emerged based on the mathematical frameworkof partially observable Markov decision processes (POMDP) [8]. This approach assumes that dialogue evolves as Markov decision process (MDP), i.e., starting in some initial state s_0, each subsequent state s_t is modelled by a transition probability. The state s_t cannot directly observe the underlying state, and our case this reflects the uncertainty in the user intention. Instead, it must maintain a probability distribution over the set of possible states, based on observations derived from spoken interaction. In this paper, the research would like to implement POMDP to our coffee maker robot. The robot appearance and its interaction with the user can be seen in Fig 1.

First International Workshop on Pattern Recognition, edited by Xudong Jiang,
Guojian Chen, Genci Capi, Chiharu Ishii, Proc. of SPIE Vol. 10011, 1001110
© 2016 SPIE · CCC code: 0277-786X/16/$18 · doi: 10.1117/12.2243589

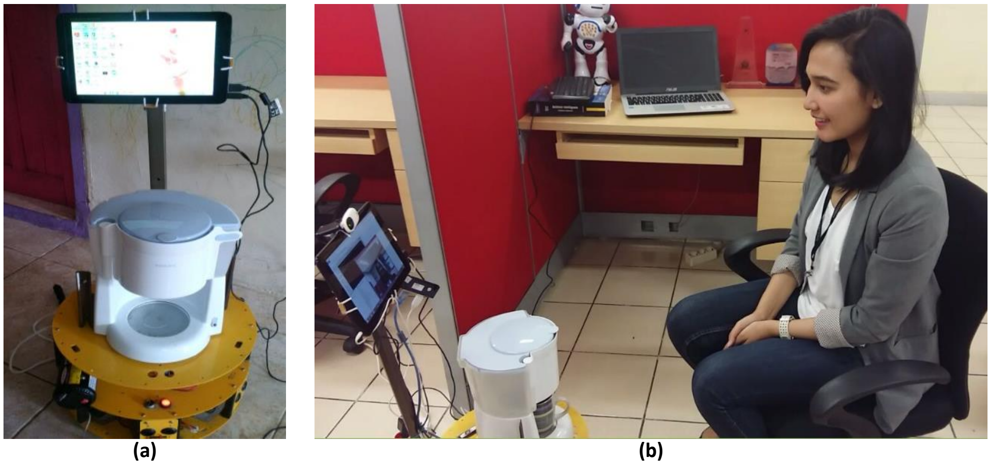


Figure 1.(a) Our appearance of coffee maker robot named CoffeBot ver 2 (b) Interaction between user and robot

2 SPOKEN DIALOGUE SYSTEM FOR COFFEE MAKER ROBOT

2.1 Architecture of Coffee Maker Robot

The coffee maker service robot is intended to be able detect human face and recognize the face of people in front to make interaction with the user. To communivate naturally, the robot should have speech recognition capability. Thus our robot is designed with input both from camera and microphone as shown in figure 2, besides the distance sensors using ultrasound. Window Tablet is used for image and speech processing. Furthermore, Arduino+Motor controller, relays and set of Coffee Maker is used as actuator.

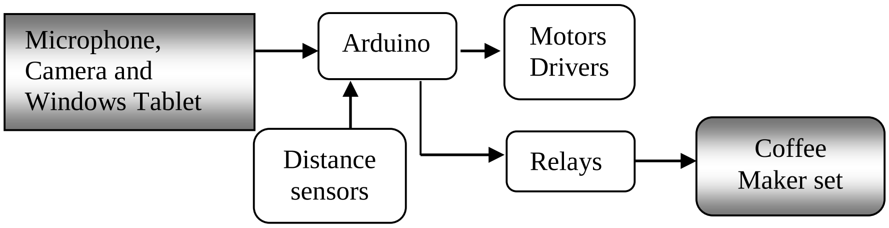


Figure 2. Architecture of our coffe maker robot

Based on our previous research [6], we use Principal Component Analysis (PCA) for face recognition and Google Speech Recognition API for speech recognition. Furthermore, FastICA is used for audio filtering and separation in noisy environment. Our main problem in here is to determine the user intention about how sweetness the ordered coffee through human-robot spoken interaction. In this paper, this spoken dialog problem is treated as a partially observable Markov decision process (POMDP). We show how this formulation establish a promising framework by empirical results. The POMDP framework commonly is used to model a variety of real-world sequential decision processes. Applications include robot navigation problems, machine maintenance, planning under uncertainty and spoken dialog system [8,9]. In the end of the paper, the dialog simulations are presented which demonstrate significant quantitative outcome.

2.2 Spoken Dialog Systems

Spoken dialog systems are machines which interact with people using spoken language. A task-oriented spoken dialog system speaks as well as understands natural language to complete a well-defined objective. This is a relatively new research area, but many task-oriented spoken dialog systems are already well advanced. Examples include a complex travel planning system, a publicly available worldwide weather information system, and an automatic call routing system [10]. Williams et al [7] have first applied Partially Observable Markov Decision Processes (POMDP) to dialog management problems. The elements of a SDS are described in Fig 3(a). In every cycle, each spoken input is converted to an abstract semantic representation in spoken language understanding (SLU) component. In dialog manager part, the system updates its internal state st and determines the next system act via a decision rule called as a policy. The outcome is then converted back into speech using a natural language generation (NLG) component.

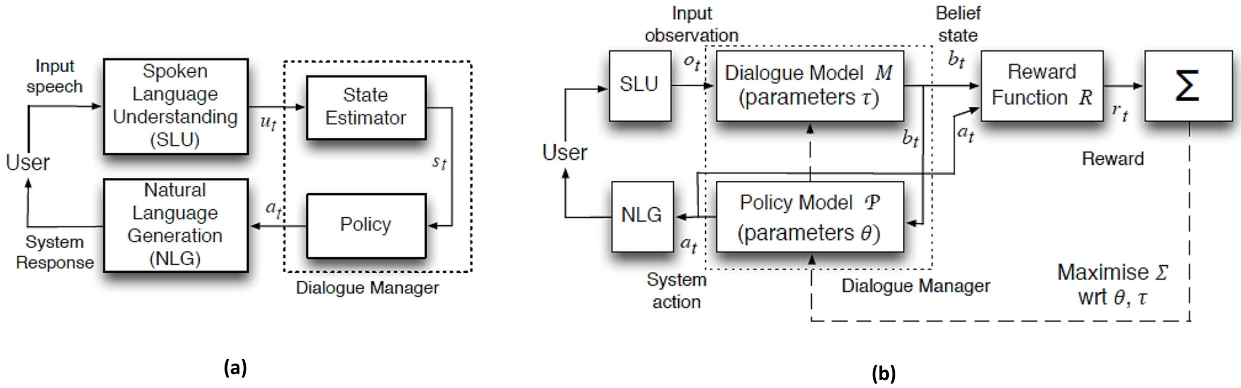


Figure 3. (a) Elements of a spoken dialogue system, (b) Components of a POMDP-based spoken dialogue system [8]

2.3 Partially Observable Markov Decision Processes (POMDP)

The partially observable Markov decision process (POMDP) is a powerful formalism for representing sequential decision problems for robot that must act under uncertainty. The algoritm address both the uncertainty in measurement and the uncertainty in control effects. Partial observability implies that the robot has to estimate a posterior distribution over possible world states [9]. At each discrete time step, the robot receives some stochastic observation related to the state of the environment,and also a special reward signal. Based on this information, the robot can execute actions to probabilistically change the state of the environment. The general goal is then to maximize the overall reward. POMDP [8] can be formally described as a tuple $\langle S,A,T,R,\Omega,O\rangle$, where:

- S is a finite set of states of the environment: $s_1, \dots, s_t \in S$
- A is a finite set of actions: $a_1, \dots, a_t \in A$
- $T : S \times A \rightarrow \Delta(S)$ is the state-transition function, giving a distribution over states of the environment, given a starting state and an action performed by the robot;
- R reward for each state/action pair: $r(s_t, a_t)$
- Ω is a finite set of noisy observations the robot can experience; and
- $O : S \times A \rightarrow \Delta(\Omega)$ is the observation function, giving a distribution over possible observations, given a starting state and an action performed by the robot.
- Probabilistic state-action transitions: $p(s_t|s_{t-1}, a_{t-1})$
- Conditional observation probabilities: $p(o_t|s_t)$

Note that the sub-tuple $\langle S,A,T,R\rangle$ represents the underlying Markov Decision process (MDP). If the observation function were to give the hidden state of the environment with perfect certainty, the problem reduces to a fully observable MDP. Components of a POMDP-based spoken dialogue system can be seen in Fig. 3(b), the input speech is regarded as a noisy observation o_t and the output is an action a_t come from policy model. Based on POMDP framework, we propose an algorithm for the Coffee Maker Service Robot as shown in algorithm 1:

Algorithm 1: Algorithm of the Coffee Maker Service Robot

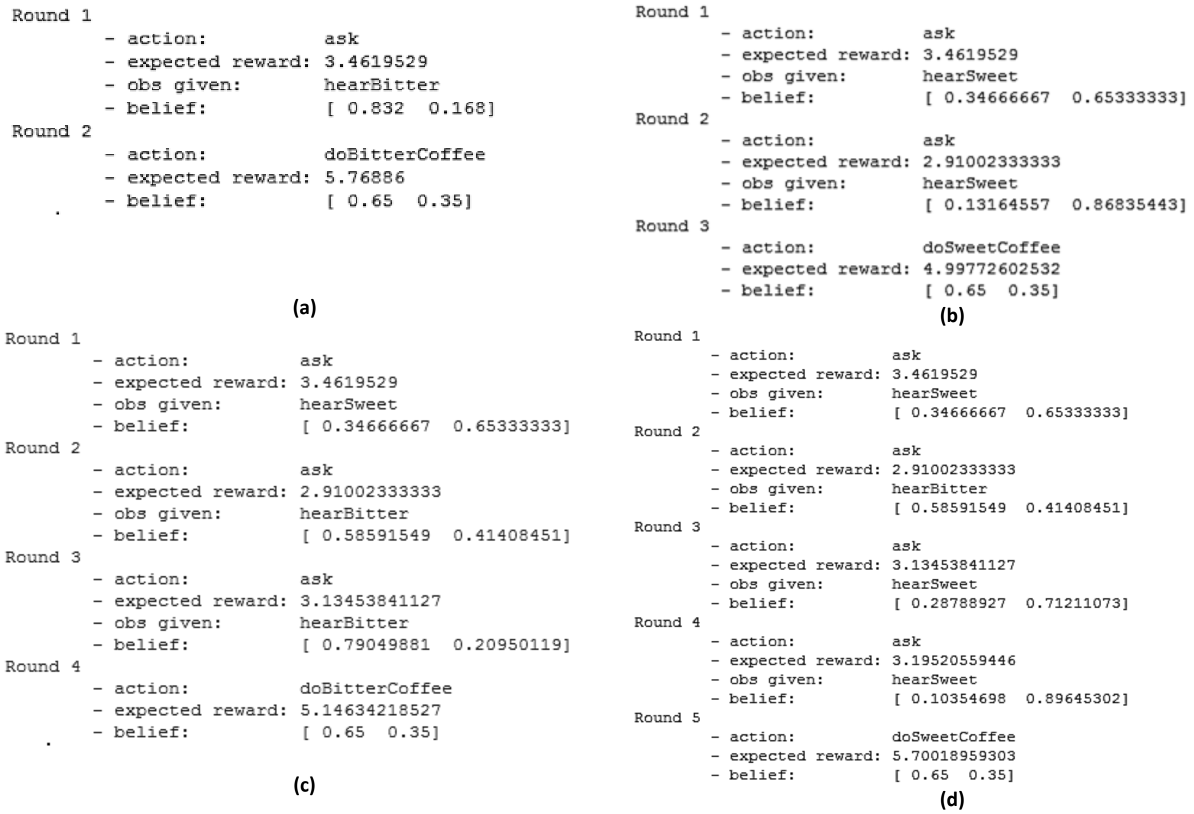
```
Get input image from the camera
Detect and recognize face using PCA
If face detected then
    Do
        If distance >=30 cm then
            Go near to the user
        endif
        Get audio from user
            Process filtering using fastICA
            Recognize each speech using Google Speech Recognition API
```

```
                Execute spoken dialog system based on POMDP
            Activate relays for making a cup of cofee based on the outcome of user’s intention
            If finished then
                robot back to the homebase and standby
            endif
        loop
else
    Robot only standby
endif
```

3 EXPERIMENTAL RESULTS

In our simulation, the coffee maker robot has only three available actions: first ask what the user wishes to do in order to infer his intention, the others doBitterCoffee and doSweetCoffe. When the user responds to a question, it is decoded as either the observation bitter or sweet. However, because of noisy environment and imprecise user’s response, these observations cannot be used to deduce the user’s intent with certainty. Based on some assumptions, if the user says sweet then an error may occur with probability 0.2, whereas if the user says bitter then an error may occur with probability 0.3. Finally, since the user wants bitter more often than sweet, the initial belief state is set to indicate the prior (0.65, 0.35), and it is reset to this value after each doSweetCoffe or doBitterCoffee action via the transition function.

The robot is designed to receive a large positive reward (+5) for getting the user’s goal correct, a very large negative reward (-20) for taking the action doSweetCoffe when the user wanted bitter (since the coffee cannot reverse back to be bitter manually), and a smaller but still significant negative reward (-10) for taking the action doBitterCoffee when the user wanted sweet (since the user can always make the coffee sweet manually). There is also a small negative reward for taking the ask action (-1), since the robot should force to achieve to its goal as quickly as possible. Based on this simulation scenario, there is only four different conversation when the spoken dialog cycle is limited to only five rounds as described in Fig 4:



```
Round 1
        - action:          ask
        - expected reward: 3.4619529
        - obs given:       hearBitter
        - belief:          [ 0.832  0.168]
Round 2
        - action:          doBitterCoffee
        - expected reward: 5.76886
        - belief:          [ 0.65  0.35]
```

(a)

```
Round 1
        - action:          ask
        - expected reward: 3.4619529
        - obs given:       hearSweet
        - belief:          [ 0.34666667  0.65333333]
Round 2
        - action:          ask
        - expected reward: 2.91002333333
        - obs given:       hearSweet
        - belief:          [ 0.13164557  0.86835443]
Round 3
        - action:          doSweetCoffee
        - expected reward: 4.99772602532
        - belief:          [ 0.65  0.35]
```

(b)

```
Round 1
        - action:          ask
        - expected reward: 3.4619529
        - obs given:       hearSweet
        - belief:          [ 0.34666667  0.65333333]
Round 2
        - action:          ask
        - expected reward: 2.91002333333
        - obs given:       hearBitter
        - belief:          [ 0.58591549  0.41408451]
Round 3
        - action:          ask
        - expected reward: 3.13453841127
        - obs given:       hearBitter
        - belief:          [ 0.79049881  0.20950119]
Round 4
        - action:          doBitterCoffee
        - expected reward: 5.14634218527
        - belief:          [ 0.65  0.35]
```

(c)

```
Round 1
        - action:          ask
        - expected reward: 3.4619529
        - obs given:       hearSweet
        - belief:          [ 0.34666667  0.65333333]
Round 2
        - action:          ask
        - expected reward: 2.91002333333
        - obs given:       hearBitter
        - belief:          [ 0.58591549  0.41408451]
Round 3
        - action:          ask
        - expected reward: 3.13453841127
        - obs given:       hearSweet
        - belief:          [ 0.28788927  0.71211073]
Round 4
        - action:          ask
        - expected reward: 3.19520559446
        - obs given:       hearSweet
        - belief:          [ 0.10354698  0.89645302]
Round 5
        - action:          doSweetCoffee
        - expected reward: 5.70018959303
        - belief:          [ 0.65  0.35]
```

(d)

Figure 4. (a) if first observation is bitter, (b) if first and second observations are sweet, (c) and (d) other variations when first observation is sweet and the second is bitter

The simulation results show the robot hear bitter, it do not make confirmation again to make sure the user intention. This behaviour is due to our initial belief is tend to doBitterCoofee decision. On the other hand, when the robot hear sweet, it have to make one more clarification about the user intention. This characteristics is come from the high punisment (-20) when the robot choose the wrong decision: taking the action doSweetCoffe when the user wanted bitter. The other variations is generated because of the unclear user intention as considered in switch responds from sweet to bitter observations. The results in this simulation is correspond with our common sense.

4 CONCLUSION

In this paper, we introduced our coffee maker service robot to deal with natural spoken conversation. The robot is based on Windows tablet which have ability for speech and image processing. Based on our previous reasearch, the speech recognition accuracy is about 86% using fastICA and face recognition accuracy about 93%. Furthermore, the current research would like to determine the user intention about how sweetness the ordered coffee through human-robot spoken interaction using POMDP. Based on simulation, this formulation show a promising framework comparing to empirical results. For the future work, our coffee maker robot will be challenged with more complex spoken interaction due to more variation of actions to do.

REFERENCES

[1] IFR, Service Robots, http://www.ifr.org/service-robots/, (2012)

[2] S. Pieska, M. Luimula, J. Jauhiainen, V. Spiz, “Social Service Robots in Wellness and Restaurant Applications”, Journal of Communication and Computer vol 10, pp. 116-123. (2013).

[3] B. Graf, C. Parlitz, M. Hagele, Robotic Home Assistant Care-O-bot® 3 Product Vision and Innovation Platform, Human-Computer Interaction, Part II, HCII 2009, LNCS 5611, Springer-Verlag, Berlin Heidelberg, (2009).

[4] Future Robot, Restaurant Service Robot FURO-R, http://www.futurerobot.com/contents_eng/sub42.htm, (2010).

[5] Widodo Budiharto, Alexander A S Gunawan, Heri Ngarianto, “Designing of Humanoid Robot with Voice Recognition Capability”, ICISIP 2015, Fukuoka - Japan.

[6] K. Nakadai, H. Nakajima, Y. Hasegawa, H. Tsujino, “Sound source separation of moving speakers for robot audition”,IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3685 - 3688. (2009). DOI: 10.1109/ICASSP.2009.4960426.

[7] Jason D. Williams, Steve Young, “Partially observable Markov decision processes for spoken dialog systems”, Journal Computer Speech and Language, Vol 21 (2), pp 393-422, (2007)

[8] Steve Young. Milica Gašić, Blaise Thomson , Jason D. Williams, “POMDP-Based Statistical Spoken Dialog Systems: A Review”, Proceedings of the IEEE Vol 101 (5), pp. 1160 – 1179, (2013)

[9] Dieter Fox, Sebastian Thrun, Wolfram Burgard, “Probabilistic Robotics”, MIT Press, (2006).

[10] A.L. Gorin, G. Riccardi and J.H. Wright., “How may I Help You?”, Speech Communication Vol 23 pp. 113-127, (1997)