*ICIC Express Letters*
Volume **11**, Number **4**, April **2017**

 ISSN 1881-803X
pp. 757–762

# COFFEE MAKER ROBOT BASED ON SIMPLE VOCABULARY AND PARTIALLY OBSERVABLE MARKOV DECISION PROCESS (POMDP)

Widodo Budiharto[1] and Chiharu Ishii[2]

[1]School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Jakarta Barat 11480, Indonesia
wbudiharto@binus.edu

[2]Department of Mechanical Engineering
Hosei University
Tokyo 184-8584, Japan
c-ishii@hosei.ac.jp

Received September 2016; accepted December 2016

Abstract. *Many recent developments of assistive robotics with face and speech recognition systems have been seen in order to interact with users naturally. In this research, we would like to propose Intelligent Coffee Maker Robot with the speech recognition based on Indonesian language using small vocabulary and statistical dialogue systems. This kind of robot can be used in the office, supermarket or restaurant. In our scenario, robots will recognize the user's face and then accept commands from the user to do an action, specifically in making 3 types of coffee, espresso, double espresso and latte, where the answer from the user will be processed to the Google Translator and the result will be compared with simple vocabulary in Indonesian language. The main problem here is to know the intention of users about how sweetness of the coffee. The intelligent coffee maker robot should conclude the user's 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 establishes a promising framework by empirical results. The dialog simulations are presented which demonstrate significant quantitative outcomes.*
**Keywords:** Speech recognition, POMDP, Coffee maker, Vocabulary, Statistical dialogue systems

1. **Introduction.** Recently, service robots and assistive robotics can be found in many human environments, in public and private. Therefore, research on assistive robotics has gained growing interest especially focusing on how to improve the interaction between robots and human beings (Human Robot Interaction). International Federation of Robots (IFR) defines service robot as a robot which operates semi- or fully autonomously to perform services useful to humans in daily life, excluding manufacturing operation [1]. Many researches of service robots and assistive robotics in different application fields have been conducted such as assistive robotics for the elderly [2], robotic home assistant [3], and robot waiters in restaurant [1,2].

Spoken dialog systems (SDS) help people accomplish a task using spoken language. Building SDSs is a challenging engineering problem in large part because automatic speech recognition (ASR) and understanding technology are error-prone [7]. Thus, the main problem here is to know the user's intention naturally through spoken interaction. This problem is commonly called as spoken dialog systems (SDS). SDS allow users to interact with others to know her intention using speech as the primary communication medium [7]. Nowadays, the common use of speech interfaces in smartphones demonstrates the value of integrating natural speech interactions into mobile applications. During the last few

years, a new approach to dialogue management has emerged based on the mathematical framework of partially observable Markov decision process (POMDP). 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 modeled by a transition probability.

2. **Problem Statement and Preliminaries.** The coffee maker service robot is intended to be able to detect human face and recognize the face of people in front to make interaction with the user. To communicate naturally, the robot should have speech recognition capability. Thus our system is designed with input both from camera and microphone from the table as shown in Figure 1. We improve our previous research [5] such as the software and Tablet PC with onboard camera used for image and speech processing. Furthermore, Arduino and relays for controlling the Nespresso machine used in this research are shown in Figure 2.

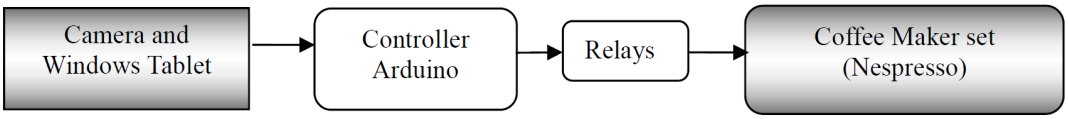



FIGURE 1. Architecture of our coffee maker robot

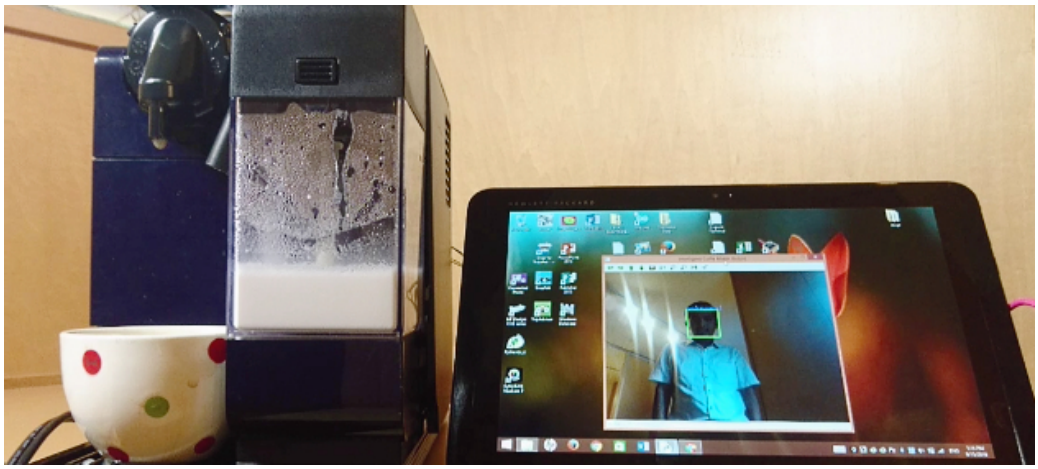


FIGURE 2. Nespresso coffee maker set and the tablet. The program for face and speech recognition system developed using Python and OpenCV [13].

**List of Vocabulary from Users**

We propose a small vocabulary for the possibility of words that will be used by users for answering the coffee maker, such as types of coffee and answer yes/no. The translation in our system is based on the Google Technology [12], and the demo of our systems is shown at [13]. The list is based on the experience when taking the data from the experiments as shown in Table 1.

TABLE 1. Proposed list of vocabulary from users based on the experience

| **Espresso** | **Double Espresso** | **Latte** | **Yes/No** |
|---|---|---|---|
| Espress | Double | Latte | Ya |
| Espresso | Double Espresso | Late | Yes |
| Istri | Double Espress | Latihan | Tidak |
| Espressi | | Lat | No |

**Spoken Dialog Systems and POMDP**

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 and Young [7] have first applied partially observable Markov decision process (POMDP) to dialog management problems. The elements of an SDS are described in Figure 3. In every cycle, each spoken input is converted to an abstract semantic representation in spoken language understanding (SLU) component. The outcome is then converted back into speech using a natural language generation (NLG) component.

The partially observable Markov decision process (POMDP) is a powerful formalism for representing sequential decision problems for agents that must act under uncertainty. The algorithm addresses 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 [8]. At each discrete time step, the agent receives some stochastic observation related to the state of the environment, as well as a special reward signal. Based on this information, the agent can execute actions to stochastically change the state of the environment. The goal of the agent is then to maximize the overall (and typically time-discounted) reward. POMDP [9] 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: $s_1, \ldots, s_t \in S$
- $A$ is a finite set of actions: $a_1, \ldots, a_t \in A$
- $T$: $S \times A \to \Delta(S)$ is the state-transition function, giving a distribution over states of the environment, given a starting state and an action performed by the agent
- $R$ reward for each state/action pair: $r(s_t, a_t)$
- $\Omega$ is a finite set of observations the agent can experience; and
- $O$: $S \times A \to \Delta(\Omega)$ is the observation function, giving a distribution over possible observations, given a starting state and an action performed by the agent
- *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 true (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 Figure 3, the input speech is regarded as a noisy observation $o_t$ and the output is an action coming from policy model.

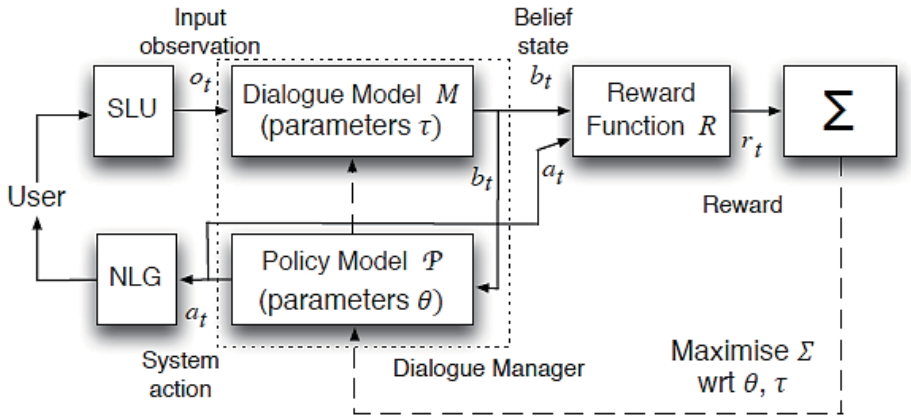



FIGURE 3. Components of a POMDP-based spoken dialogue system [8]

Based on POMDP framework, we propose an algorithm for the Coffee Maker Service Robot as shown in Algorithm 1:

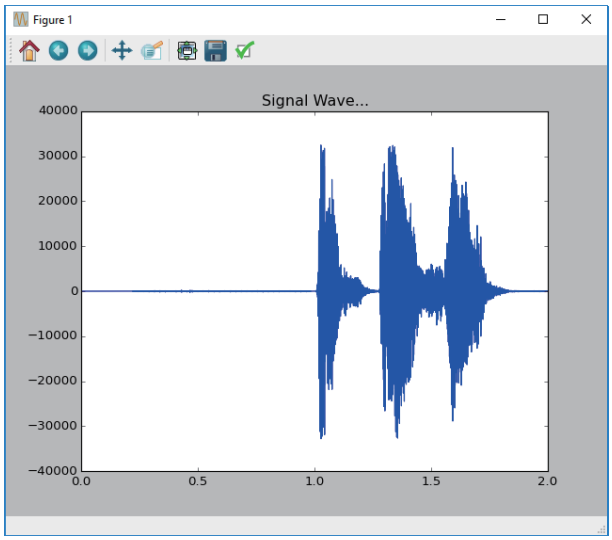
Algorithm 1. **Algorithm of the Coffee Maker Service Robot with simple vocabulary and POMDP**

```
Get input image from the camera
Detect and recognize face using PCA
If face detected then
Do
  Welcoming message and asking the user
  Get answer from user (.WAV file)
  Recognize each speech using Google Speech Recognition API
  Execute spoken dialog system based on POMDP
  Check answer based on simple vocabulary
  If answer is in simple vocabulary then
      Activate relays for making a cup of coffee based on the outcome of user's intention
      If finished then
          Coffee maker Standby
      endif
  else
      Ask again
  endif
  Coffee maker Standby
loop
else
Coffee maker standby
endif
```

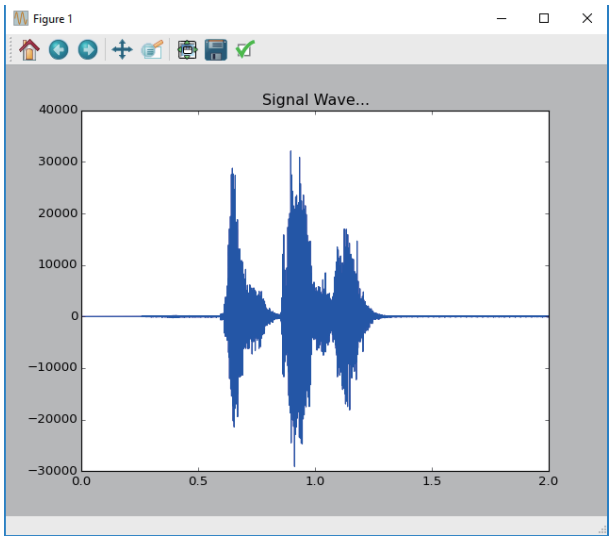
3. **Main Results.** In our experiments, the coffee maker robot has only three available actions: first ask what the user wishes to do in order to infer his intention, the others doEspresso, doDoubleEspresso and doLatte. When the user responds to a question, it is decoded as either the observation Espresso, Double Espresso or sweet. However, because of noisy environment and user's imprecise response, these observations cannot be used to deduce the user's intent with certainty. Based on some assumptions, if the user says Latte then an error may occur with probability 0.2, whereas if the user says Espresso then an error may occur with probability 0.3. Finally, since the user wants Espresso 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 doLatte or doEspresso 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 doLatte when the user wanted Espresso (since the coffee cannot reverse back to be bitter manually), and a smaller but still significant negative reward $(-10)$ for taking the action doEspresso 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 machine should force to achieve to its goal as quickly as possible.

We check the effect of the distance with noise, which shows in Figure 4(b) that the signal with noise has different patterns and many noises and will cause bad words translation.

The comparative experiment is with 10 persons (Japanese students at Hosei University, Japan) that try 5 times each command randomly. There are total 15 times for one kind of situation. The number of total commands/persons is $15 \times 3 = 45$ times. The result is shown in Table 2. From Table 2, we see that with POMDP if the machine did not get the correct answer compared with the vocabulary, it will ask again, so the average




(a)




(b)

FIGURE 4. (a) Without noise and distance between robot and a user is 50cm, and (b) with noise and the distance is 120cm

TABLE 2. Accuracy with 10 persons with and without POMDP and varying distance of users

| **Subjects** | **Without POMDP/noisy** | | **With POMDP/quiet** | | **With POMDP/noisy** | |
|---|---|---|---|---|---|---|
| | ***50cm*** | ***120cm*** | ***50cm*** | ***120cm*** | ***50cm*** | ***120cm*** |
| subject 1 | 60% | 40% | 93.3% | 86.7% | 86.7% | 80% |
| subject 2 | 46.7% | 26.7% | 80% | 73.3% | 66.7% | 46.7% |
| subject 3 | 60% | 33.3% | 100% | 93.3% | 86.7% | 66.7% |
| subject 4 | 26.7% | 20% | 86.7% | 73.3% | 60% | 60% |
| subject 5 | 33.3% | 26.7% | 93.3% | 73.3% | 60% | 60% |
| subject 6 | 26.7% | 20% | 86.7% | 53.3% | 66.7% | 33.3% |
| subject 7 | 26.7% | 6.7% | 66.7% | 60% | 73.3% | 26.7% |
| subject 8 | 66.7% | 26.7% | 93.7% | 93.3% | 93.3% | 93.3% |
| subject 9 | 33.3% | 33.3% | 86.7% | 93.3% | 86.7% | 60% |
| subject 10 | 13.3% | 13.3% | 60% | 66.7% | 40% | 53.3% |
| **Average accuracy** | **39.3%** | **24.7%** | **84.7%** | **76.7%** | **72%** | **58%** |

accuracy reaches 84.7%. However, with noise and POMDP, maybe because of the noisy environment, the system asks twice but still does not get the correct answer. The accuracy with noise and POMDP by varying the distance is shown, too, and it can be seen that the reduced signal will cause errors in words translation.

Sometimes emotion from user is important. So, next investigation is the machine that could help people in making right decisions by recognizing emotions, especially in irrational situations where decisions have to be made faster than a rational performing mind, as proposed at [11].

4. **Conclusions.** We successfully implement the speech recognition with simple vocabulary and POMDP method for intelligent coffee maker robot. The experimental results show that when the robot hears expresso, it does not make confirmation again to make sure the user intention. This behavior is due to the fact that our initial belief is doEspresso decision. On the other hand, when the robot hears sweet, it has to make one more clarification about the user intention. This characteristic comes from the high punishment ($-20$) when the robot chooses the wrong decision: taking the action doLatte when the user wanted bitter. The other variation is generated because of the unclear user intention as considered in switch respones from sweet to bitter observations. The results in this simulation correspond with our common sense. For the future work, our coffee maker robot will be challenged with more complex spoken interaction due to more variation of actions to do, and next, additional features such as recognizing emotions, especially in irrational situations where decisions have to be made faster than a rational performing mind.

**Acknowledgment.** This work is fully supported by Hosei University-Tokyo as a visiting researcher in 2016.

## REFERENCES

[1] IFR, *Service Robots*, http://www.ifr.org/service-robots/, 2012.
[2] S. Pieska, M. Luimula, J. Jauhiainen and 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 and M. Hägele, 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] W. Budiharto, Meiliana and A. A. S. Gunawan, Development of coffee maker service robot using speech and face recognition systems using POMDP, *The 1st International Workshop on Pattern Recognition*, Tokyo, Japan, 2016.
[6] K. Nakadai, H. Nakajima, Y. Hasegawa and H. Tsujino, Sound source separation of moving speakers for robot audition, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.3685-3688, 2009.
[7] J. D. Williams and S. Young, Partially observable Markov decision processes for spoken dialog systems, *Journal Computer Speech and Language*, vol.21, no.2, pp.393-422, 2007.
[8] S. Young, M. Gaši, B. Thomson and J. D. Williams, POMDP-based statistical spoken dialog systems: A review, *Proc. of the IEEE*, vol.101, no.5, pp.1160-1179, 2013.
[9] D. Fox and S. 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.
[11] S. Lugović, I. Dunder and M. Horvat, Techniques and applications of emotion recognition in speech, *Proc. of the 39th International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2016*, pp.1278-1283, 2016.
[12] *Google Translatoer*, www.translate.google.com.
[13] *Intelligent Coffee Maker Robot*, https://www.youtube.com/watch?v=cJLUmAITexI.