Conversational Strategies for Robustly Managing Dialog in Public Spaces

Aasish Pappu **Ming Sun** **Seshadri Sridharan** **Alexander I. Rudnicky**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh PA, USA
{aasish,mings,seshadrs,air}@cs.cmu.edu

Abstract

Open environments present an attention management challenge for conversational systems. We describe a kiosk system (based on Ravenclaw–Olympus) that uses simple auditory and visual information to interpret human presence and manage the system's attention. The system robustly differentiates intended interactions from unintended ones at an accuracy of 93% and provides similar task completion rates in both a quiet room and a public space.

1 Introduction

Dialog systems designers try to minimize disruptive influences by introducing physical and behavioral constraints to create predictable environments. This includes using a closed-talking microphone or limiting interaction to one user at a time. But such constraints are difficult to apply in public environments such as kiosks (Bohus and Horvitz, 2010; Foster et al., 2012; Nakashima et al., 2014), in-car assistants (Kun et al., 2007; Hofmann et al., 2013; Misu et al., 2013) or on mobile robots (Haasch et al., 2004; Sabanovic et al., 2006; Kollar et al., 2012). To implement dialog systems that operate in public spaces, we have to relax some of these constraints and deal with additional challenges. For example, the system needs to select the correct interlocutor, who may be only one of several possible ones in the vicinity, then determine whether they are initiating the process of engaging with the system.

In this paper we focus on the problems of identifying a potential interlocutor in the environment, engaging them in conversation and providing suitable channel-maintenance cues (Bruce et al., 2002; Fukuda et al., 2002; Al Moubayed and Skantze, 2011). We address these problems in the context of a simple application, a kiosk agent that accepts tasks such as taking a message to a named recipient. To evaluate the effectiveness of our approach we compared the system's ability to manage conversations in a quiet room and in a public area.

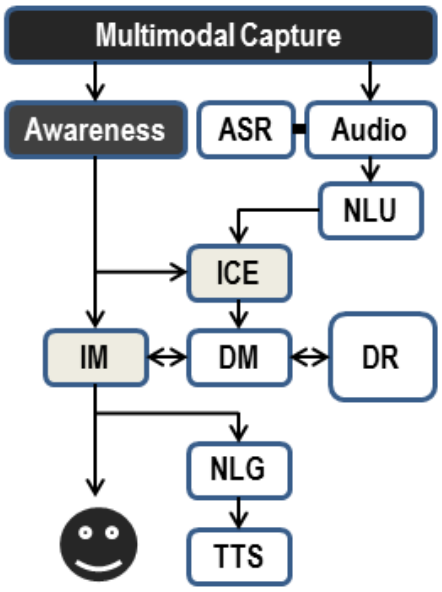

Figure 1: Ravenclaw–Olympus augmented with multimodal input and output functions.

The remainder of this paper is organized as follows: we first describe the system architecture, then present the evaluation setup and the results, then review related work and finally conclude with an analysis of the study.

2 System Architecture

Figure 1 shows the architecture; it incorporates Ravenclaw/Olympus (Bohus et al., 2007) standard components (in white), new components (in black) and modified ones (shaded). In the system pipeline, the Audio Server receives audio from a microphone, endpoints it and sends it to the ASR

Proceedings of the of the EACL 2014 Workshop on Dialogue in Motion (DM), pages 63–67,
Gothenburg, Sweden, April 26-30 2014.

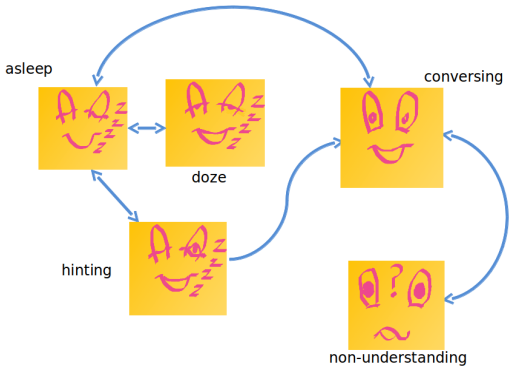

Figure 2: Face states; some are animations.

engine (PocketSphinx); the decoding is passed to NLU (Phoenix parser). ICE (Input Confidence Estimation) (Helios) assigns confidence scores for the input concepts. Based on user's input and the context, the Dialog Manager (DM) determines what to do next, perhaps using data from the Domain Reasoner (DR). An Interaction Manager (IM) initiates a spoken response using Natural Language Generation (NLG) and Text-to-Speech (TTS) component.

Three components were added: (1) *Multimodal Capture* acquires audio and human position data using a Kinect device [1]. (2) *Awareness* determines whether there is a potential interlocutor in the vicinity and their current position, using skeletal and azimuth information. (3) *Talking Head* that conveys the system's state (as shown in Figure 2): whether it's active (*conversing* and *hinting*) or idle (*asleep* and *doze*) and whether focused concepts are grounded (*conversing* and *non-understanding*); certain state representations (e.g., *conversing*) are coordinated with the TTS component.

3 Evaluation

A robust system should be able to function as well in a difficult situation as in a controlled one. We compare the system's performance in two environments, public and quiet, and evaluate the (a) system's awareness of intended users, and its (b) end-to-end performance.

The same twenty subjects participated in both experiments: a mix of American, Indian, Chinese and Hispanic with different fluency levels of English. None of them had previously interacted with this system prior to this study.

The subjects were told that they would interact with a virtual agent displayed on a screen. Their task for the awareness experiment was to make the agent aware that they wished to interact. For the end-to-end system performance, the task was to instruct the agent to send a message to a named recipient.

3.1 Situated Awareness

We define situated awareness as correctly engaging the intended interlocutor (i.e., verbally acknowledge the user's presence) under two conditions. When the user is positioned (i) inside the visual range of the Kinect at LOC-0 in Figure 3(a); and (ii) outside the visual range of the Kinect at LOC-1 in Figure 3(a). We used the effective range of the camera's documented horizontal field of view (57°); hereafter referred as its *cone-of-awareness*.

We conducted the awareness experiment in a public space, a lounge at a hub connecting multiple corridors. The area has tables and seating, self-serve coffee, a microwave oven, etc. The experiment was conducted during regular hours, between 10am to 6pm on weekdays. During these times we observed occupants discussing projects, preparing food, making coffee, etc. No direct attempt was made to influence their behavior and we believe that they made no attempt to accommodate our activities. Accordingly, the natural sound level in the room varied in unpredictable ways. To supplement naturally-occurring sounds, we played audio of a conversation between two humans, an extract from the SwitchBoard corpus (Graff et al., 2001). It was played using a loudspeaker placed at LOC-2 in Figure 3(a). The locations (0, 1, and 2) are all 1.5m from the Kinect, which we deemed to be a comfortable distance for the subjects. LOC-1 and LOC-2 are 70° to the left and right of the Kinect, outside its cone.

To detect the presence of an intended user, we build an awareness model that uses three sensory streams viz., voice activity, skeleton, and sound source azimuth. This model relies on the coincidence of azimuth angle and the skeleton angle (along with voice activity) to determine the presence of an intended user. We compare the pro-

[1] See `http://www.microsoft.com/en-us/Kinectforwindows/develop/`. Three sources are tapped: the beam-formed audio, the sound source azimuth and skeleton coordinates. Video data are not used.

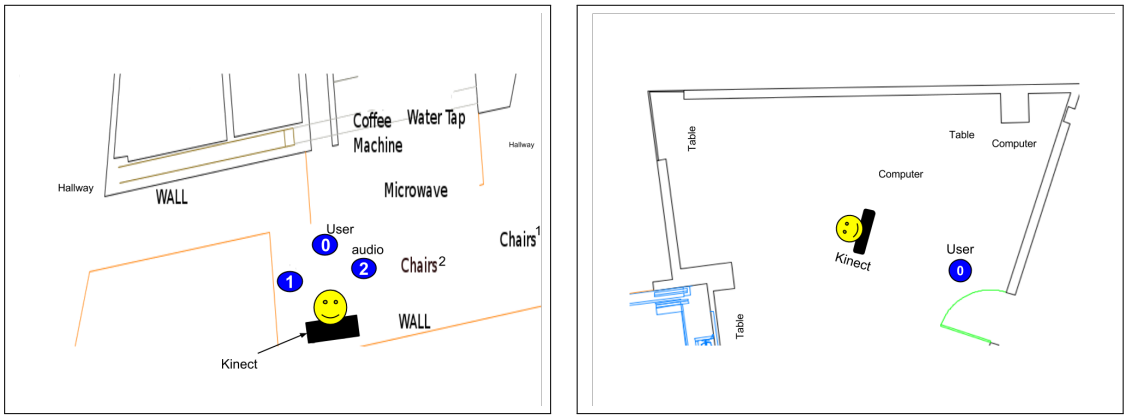

Figure 3: (a) Plan of Public Space (lounge);(b) Plan of Quiet Room (lab). Dark circled markers indicate locations (LOC-0, LOC-1, LOC-2), discussed in the text.

Condition	Voice	+Skeleton	+Azimuth
Outside the cone	28%	–	**93%**
Inside the cone	–	25%	**93%**

Table 1: Accuracy for the Awareness Detection

posed model with two baselines: (1) conventional voice-activity-detection (VAD): once speech is detected the system responds as if a conversation is initiated and (2) based on skeleton plus VAD: once the skeleton appears in front of the Kinect and a voice is heard, the system engages in conversation.

Table 1 shows the combination of sensory streams we used under two conditions. For the outside-the-cone condition, the participants stand in LOC-1 as shown in Figure 3(a) and follow the instructions from the agent. Initially, the subject's skeleton is invisible to the agent; however the subject is audible to the agent. Therefore, in certain combinations of sensors (e.g., `voice + skeleton` model and `voice + skeleton + azimuth` model) the system attempts to guide them to move in front of it, i.e. to LOC-0, an ideal position for interacting with the system. For inside-the-cone condition, subjects stand at LOC-0 where the agent can sense their skeleton.

When user stands at LOC-1 i.e., outside-the-cone `voice + skeleton model` and `voice + skeleton + azimuth` models are functionally the same since the source of distraction has no skeleton in the cone. When user stands at LOC-0, i.e., inside-the-cone `voice` alone is the same as `voice + skeleton` model since the agent always sees a skeleton in front of it. Therefore, this variant was not used.

We treated awareness detection as a binary decision. An utterance is classified either as "intended" or "unintended". We manually labeled the utterances whether they were directed at the system ("intended"), "unintended" otherwise. Accuracy on "intended" speech is reported in the Table 1. Within each condition, the order of the experiments with different awareness strategies was randomized.

We observe that the `voice + skeleton + azimuth` model proves to be robust in the public space. Its performance is significantly better, $t(38) = 8.1$, $p \approx 0.001$, compared to the other baselines in both conditions. This result agrees with previous research (Haasch et al., 2004; Bohus and Horvitz, 2009) showing that a fusion of multimodal features improves performance over a unimodal approach. Our result indicates that a simple heuristic approach, using minimal visual and audio features, provides usable attention management in open environments. This approach helped the system handle a complex interaction scenario such as out-of-cone speech directed to the system. If the speaker is out of range but is producing possibly system-directed utterances, system urges them to step to the front. We believe it can be extended to other complex cases by introducing additional logic.

3.2 End-to-End System Performance

To investigate the effect of the environment, we compare the system's performance in public space and quiet room. The average noise level in the quiet room is about 47dB(A) with computers as

Metric	Public Space	Quiet Room
Success Ratio	15/20	16/20
Avg # Turns	14.2	16.4
Concept Acc	67%	68%

Table 2: Public Space vs Quiet Room Performance

the primary source of noise. The background sound level in the public space was 46dB; other natural sources ranged up to 57dB. The audio distractor measured 57dB. The same ASR acoustic models and processing parameters were used in both environments. The participant stood at LOC-0 in Figure 3(a) during the public space experiment and Figure 3(b) during the quiet room experiment. In both experiments, LOC-0 is 1.5m away from the system. We used the `voice + skeleton + azimuth` model to discriminate user speech from distractions in the environment.

We gave each participant a randomized series of message-sending tasks, e.g. "send a message to ⟨person⟩ who is in room ⟨number⟩". Subjects had a maximum of 3 minutes to complete; each task required 7 turns. The number of tasks completed (over the group) is reported in terms of task "success-ratio". Table 2 shows the success-ratio of the task, the average number of turns needed to complete the task, and the system's per-utterance concept accuracy (Boros et al., 1996). There were no statistically significant differences between quiet room and public space, ($t(38) < 2, p > 0.5$, on any metric). We conclude that the channel maintenance technique we tested was equally effective in both environments.

4 Related Work

The problem of deploying social agents in public spaces has been of enduring interest; (Bohus and Horvitz, 2010) list engagement as a challenge for a physically situated agent in open-world interactions. But the problem was noted earlier and solutions were proposed; e.g a "push-to-talk" protocol to signal the onset of intended user speech (Stent et al., 1999). (Sharp et al., 1997; Hieronymus et al., 2006) described the use of attention phrase as a required prefix to each user input. Although explicit actions are effective, they need to be learned by users. This may not be practical for systems in public areas engaged by casual users.

A more robust approach involves fusing several sources of information such as audio, gaze and pose(Horvitz et al., 2003; Bohus and Horvitz, 2009) (Hosoya et al., 2009; Nakano and Ishii, 2010). Previous works have shown that fusion of different sensory information can improve attention management. The drawback of such approaches is in the complexity of the sensor equipment. Our work attempts to create the relevant capabilities using a simple sensing device and relying on explicitly modeled conversational strategies. Others are also using the Microsoft Kinect device for research in dialog. For example, (Skantze and Al Moubayed, 2012) and (Foster et al., 2012) presented a multiparty interaction systems that use Kinect for face tracking and skeleton tracking combined with speech recognition.

In our current work, we show that situational awareness can be integrated into an existing dialog framework, Ravenclaw–Olympus, that was not originally designed with this functionality in mind. The source code of the framework presented in this work is publicly available for download [1] and the acoustic models that have been adapted to the Kinect audio channel [2]

5 Conclusion

We found that a conventional spoken dialog system can be adapted to a public space with minimal modifications to accommodate additional information sources. Investigating the effectiveness of different awareness strategies, we found that a simple heuristic approach that uses a combination of sensory streams viz., voice, skeleton and azimuth, can reliably identify the likely interlocutor. End-to-end system performance in a public space is similar to that observed in a quiet room, indicating that, at least under the conditions we created, usable performance can be achieved. This is a useful finding. We believe that on this level, channel maintenance is a matter of articulating a model that specifies appropriate behavior in different states defined by a small number of discrete features (presence, absence, coincidence). We conjecture that such a framework is likely to be extensible to more complex situations, for example ones involving multiple humans in the environment.

[1] `http://trac.speech.cs.cmu.edu/repos/olympus/tags/KinectOly2.0/`

[2] `http://trac.speech.cs.cmu.edu/repos/olympus/tags/KinectOly2.0/Resources/DecoderConfig/AcousticModels/Semi_Kinect.cd_semi_5000/`

References

[Al Moubayed and Skantze2011] S. Al Moubayed and G. Skantze. 2011. Turn-taking control using gaze in multiparty human-computer dialogue: Effects of 2d and 3d displays. In *Proceedings of AVSP, Florence, Italy*, pages 99–102.

[Bohus and Horvitz2009] D. Bohus and E. Horvitz. 2009. Dialog in the open world: platform and applications. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 31–38. ACM.

[Bohus and Horvitz2010] D. Bohus and E. Horvitz. 2010. On the challenges and opportunities of physically situated dialog. In *2010 AAAI Fall Symposium on Dialog with Robots*. AAAI.

[Bohus et al.2007] D. Bohus, A. Raux, T.K. Harris, M. Eskenazi, and A.I. Rudnicky. 2007. Olympus: an open-source framework for conversational spoken language interface research. In *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies*, pages 32–39. Association for Computational Linguistics.

[Boros et al.1996] M. Boros, W. Eckert, F. Gallwitz, G. Gorz, G. Hanrieder, and H. Niemann. 1996. Towards understanding spontaneous speech: Word accuracy vs. concept accuracy. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 2, pages 1009–1012. IEEE.

[Bruce et al.2002] A. Bruce, I. Nourbakhsh, and R. Simmons. 2002. The role of expressiveness and attention in human-robot interaction. In *Proceedings of 2002 IEEE International Conference on Robotics and Automation*, volume 4, pages 4138–4142. IEEE.

[Foster et al.2012] M.E. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R.P.A. Petrick. 2012. "two people walk into a bar": Dynamic multi-party social interaction with a robot agent. In *Proc. of the 14th ACM International Conference on Multimodal Interaction ICMI*.

[Fukuda et al.2002] T. Fukuda, J. Taguri, F. Arai, M. Nakashima, D. Tachibana, and Y. Hasegawa. 2002. Facial expression of robot face for human-robot mutual communication. In *Proceedings of 2002 IEEE International Conference on Robotics and Automation*, volume 1, pages 46–51. IEEE.

[Graff et al.2001] D. Graff, K. Walker, and D. Miller. 2001. Switchboard cellular part 1 transcribed audio. In *Linguistic Data Consortium, Philadelphia*.

[Haasch et al.2004] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, GA Fink, J. Fritsch, B. Wrede, and G. Sagerer. 2004. Biron–the bielefeld robot companion. In *Proc. Int. Workshop on Advances in Service Robotics*, pages 27–32. Stuttgart, Germany: Fraunhofer IRB Verlag.

[Hieronymus et al.2006] J. Hieronymus, G. Aist, and J. Dowding. 2006. Open microphone speech understanding: correct discrimination of in domain speech. In *Proceedings of 2006 IEEE international conference on acoustics, speech, and signal processing*, volume 1. IEEE.

[Hofmann et al.2013] H. Hofmann, U. Ehrlich, A. Berton, A. Mahr, R. Math, and C. Müller. 2013. Evaluation of speech dialog strategies for internet applications in the car. In *Proceedings of the SIGDIAL 2013 Conference*, pages 233–241, Metz, France, August. Association for Computational Linguistics.

[Horvitz et al.2003] E. Horvitz, C. Kadie, T. Paek, and D. Hovel. 2003. Models of attention in computing and communication: from principles to application. In *Communications of the ACM*, volume 46, pages 52–59.

[Hosoya et al.2009] K. Hosoya, T. Ogawa, and T. Kobayashi. 2009. Robot auditory system using head-mounted square microphone array. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 2736–2741. IEEE.

[Kollar et al.2012] T. Kollar, A. Vedantham, C. Sobel, C. Chang, V. Perera, and M. Veloso. 2012. A multi-modal approach for natural human-robot interaction. In *Proceedings of 2012 International Conference on Social Robots*.

[Kun et al.2007] A. Kun, T. Paek, and Z. Medenica. 2007. The effect of speech interface accuracy on driving performance. In *INTERSPEECH*, pages 1326–1329.

[Misu et al.2013] T. Misu, A. Raux, I. Lane, J. Devassy, and R. Gupta. 2013. Situated multi-modal dialog system in vehicles. In *Proceedings of the 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction*, pages 25–28. ACM.

[Nakano and Ishii2010] Y. Nakano and R. Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 139–148. ACM.

[Nakashima et al.2014] Taichi Nakashima, Kazunori Komatani, and Satoshi Sato. 2014. Integration of multiple sound source localization results for speaker identification in multiparty dialogue system. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 153–165. Springer New York.

[Sabanovic et al.2006] S. Sabanovic, M.P. Michalowski, and R. Simmons. 2006. Robots in the wild: Observing human-robot social interaction outside the lab. In *Advanced Motion Control, 2006. 9th IEEE International Workshop on*, pages 596–601. IEEE.

[Sharp et al.1997] R.D. Sharp, E. Bocchieri, C. Castillo, S. Parthasarathy, C. Rath, M. Riley, and J. Rowland. 1997. The watson speech recognition engine. In *Proceedings of 1997 IEEE international conference on acoustics, speech, and signal processing*, volume 5, pages 4065–4068. IEEE.

[Skantze and Al Moubayed2012] G. Skantze and S. Al Moubayed. 2012. Iristk: a statechart-based toolkit for multi-party face-to-face interaction. In *Proc. of the 14th ACM International Conference on Multimodal Interaction ICMI*.

[Stent et al.1999] A. Stent, J. Dowding, J. Gawron, E. Bratt, and R. Moore. 1999. The commandtalk spoken dialogue system. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 183–190. ACL.