2009-10-01

A Mobile Multimodal Dialogue System for Location Based Services

Niels Schütte
Technological University Dublin, niels.schutte@student.die.ie

John D. Kelleher
Technological University Dublin, john.d.kelleher@tudublin.ie

Brian MacNamee
Technological University Dublin, brian.macnamee@tudublin.ie

A Mobile Multimodal Dialogue System for Location Based Services

Niels Schütte, John Kelleher, Brian Mac Namee

DIT Dublin, Aungier Street, Dublin 6

Niels.Schutte@student.dit.ie, John.Kelleher@comp.dit.ie, Brian.MacNamee@comp.dit.ie

Abstract

This paper describes ongoing work on the dialogue management components for LOK8, a multimodal dialogue system. We describe our plans for a basic architecture of the system, the rough modules and outline the kinds of models in the project, as well as the next steps in our work.

Keywords: Dialogue Systems, Multimodal Interaction, Location Based Services

1 Introduction

The goal of the LOK8 project is to develop a mobile multimodal dialogue system that allows users to access **Location based Services** (LBS) using a mobile device like an iPhone or a Google Android phone. LBS may offer the user functionality such as supplying information about nearby objects, or giving navigation help. We believe that such a system may benefit from having a natural language interface since natural language allows intuitive interaction, and it also removes the need for graphical interfaces which can be difficult to operate on a small mobile device. The use of pointing gestures as input and output modalities will also compensate the difficulties of expressing spatial concepts in language. It has been shown that spatial domains are especially well suited for multimodal interaction ([Oviatt, 1997]).

One possible example application scenario for the LOK8 system is a museum, where the users may walk around with their device and point the device at exhibits to request descriptions from the system, or ask specific questions. On the other hand the system may point out if the user is approach interesting exhibits or give directions.

This paper describes the work that is planned for the dialogue aspects of the LOK8 system. We describe the general architecture of the dialogue related parts of the system and the different components and models in that architecture.

2 The system

The LOK8 system (figure 1) allows interaction in a number of modalities. The primary modality is the use of spoken natural language. The user can engage in natural conversation with the system to access a number of different LBS. On the other hand the system may, depending on its configuration, take initiative and initiate conversation, for example by calling attention to points of interest the user is passing or offering services that are available in the current context.

In conjunction with **Global Localization** the user will be able to use the mobile device to perform gestures such as pointing towards objects in the environment or performing other certain gestures, such as shaking the device as a refusing gesture.

The main output modality will again be natural language. We will employ an 3D animated virtual agent as a primary point of interaction with the user i.e. the actions of the systems will primarily be

expressed as actions of the agent and user action will be directed towards the agent. In addition to that we are planning to allow the agent to have a certain amount of personality to make interaction more interesting.

This agent will in normal operation be displayed on the display of the device. We are also planning to prepare special environments in which the system has access to devices in the environment such as large projection displays or special loudspeaker systems. Apart from just displaying the agent, the display may also be used to display general information such as lists of data or maps.

The use of displays on modern mobile devices also opens up the possibility of using touch displays as an additional input modality. However, interaction with speech and gestures is our primary research focus.

Apart from the strictly dialogic interaction with the agent we are also pursuing methods of sonification to convey events and states of the system by using appropriate sounds or musical cues. It will be an interesting research question to investigate in how far it is beneficial to integrate this modality with the classical idea of dialogue, or if it is preferable to utilize it to specifically communicate content that may be hard to express in verbal dialogue such as switching to special operation modes.

The remainder of the paper is organized as follows: In section 3 we describe the proposed architecture of the system and sketch the different modules. Then we describe the different models that are going to be necessary to represent data in the system in section 4. Finally, we are going to give a short overview of the next steps in our works in section 6.

3 Architecture

Our current architecture (figure 2) is based on a basic pipeline approach that unifies inputs from different modalities, processes them in a dialogue model and produces output that is then distributed over the different output modalities depending on the current situational context. It is in certain respects similar to the architecture view presented in [Herzog and Reithinger, 2006] for the SmartKom system. There the authors present a flexible architecture for multimodal dialogue systems for different mobile or static application scenarios. Especially the "Mobile Travel Companion" scenario is in some respects similar to the functionality considere for LOK8. However we think that because of the stronger interaction with the physical environment in LOK8, there be sufficiently new research questions.

Speech is picked up by a microphone. A **Speech Recognition** component extracts hypotheses about the content of the utterance. The result is run through a **Language Interpretation** component that generates a specification of the content of the utterance. This result is entered into a **Modality Fusion** component that feeds the unified event into the **Dialogue Manager**.

In parallel the sensors of the device pick up movement and position of the device. This data is analyzed in terms of a gesture vocabulary by a **Gesture Recognizer**. We distinguish between two classes of movements: pointing gestures and general movement gestures. Pointing gestures are interpreted as deictic references to objects in the environments. Movement gestures are gestures that involve some general movement of the device such as shaking, waving or drawing lines. They are interpreted depending on the state of the dialogue. A linear movement of the device may be for example be interpreted as describing a direction or the size of an object.

The **Dialogue Management** component produces a reaction, taking into account the different context models (see section 4). This reaction is split up into concrete actions in the output modalities in the **Modality Distribution** module. This module decides what parts of the action to express in each available output modality and feeds the information into the respective generation module, each of which then generates the surface action that is presented to the user.

The model is not a strict pipeline model since we plan to make the system capable of taking initiative and producing actions without requests by the user. These actions may be for example be caused by events in the environment such as new services becoming available because of some change of conditions.

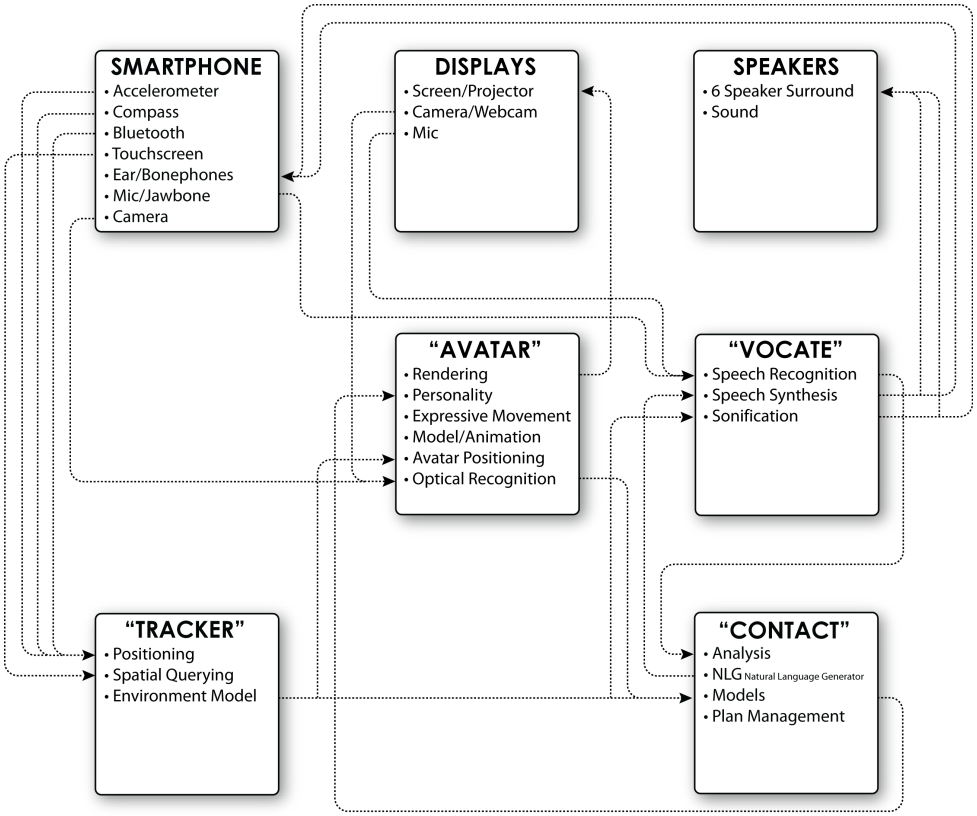

Figure 1: Architecture of the LOK8 project.

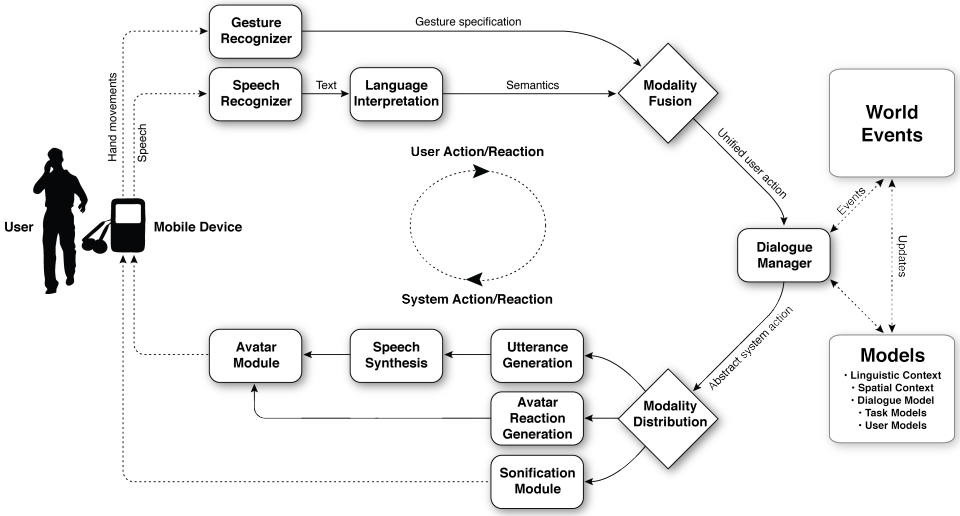

Figure 2: Architecture of the Contact Module.

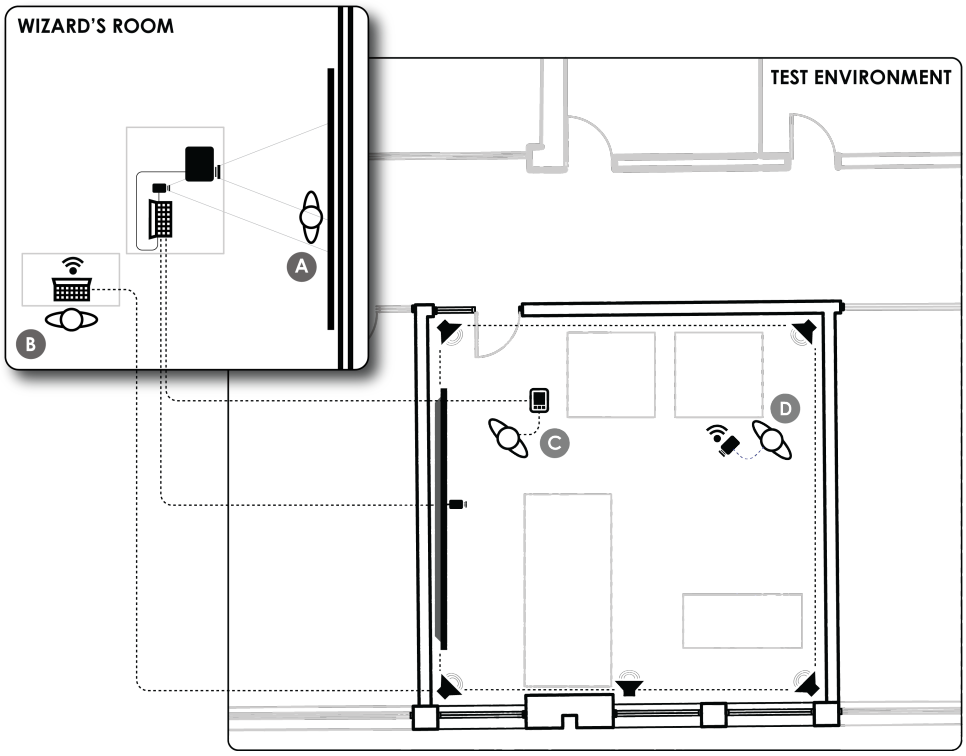

Figure 3: Setup for the Wizard of Oz experiment.

4 Models

We have to utilize and maintain a number of models and representations in the system. All information exchange interfaces between the different modules in the architecture require a form of representation that is appropriate for the communicated contents and the purposes of the modules involved. Another influence on the modeling decisions arises from the fact that modalities have to be combined. The representations should therefore use compatible formalisms.

The dialogue manager requires a **Dialogue Model** that abstractly defines the dialogues the system is capable of performing. At this point we have not made any decision as to what paradigm we are going to use for this model. This will also depend on the results of experiments described under section 6.

Apart from this model of general dialogic competence we may develop a separate model for the different specialized tasks and services that the system is to offer (**Task Models**). For these models it may be advisable to develop a formalism that allows rapid development of such task models. To develop such a formalism that takes into account the dialogic competence as well as the spatial situatedness of the system appears as an interesting challenge.

During the dialogue, the system will have to access different context models. The **Linguistic Context** model is a model of what has been discussed in the dialogue, and may be used for e.g. resolving the referents of pronouns. The **Spatial Context** model is a model of the environment in which the system is situated and is used e.g. to resolve the referents of expressions that refer to objects in the environment, such as “the blue door”. To enable such a resolution it is necessary for the model to contain information about the position and properties such as appearance of objects. Another important aspect that has to be captured by this model is the salience of objects, since this may be decisive for the resolution of expressions.

While this model is more like a representation of the perception of the environment, we also need a strictly geometric view on the environment, that has to be aligned with the actual environment the user is in. Knowledge about the position of objects enables us e.g. to resolve pointing references. Both resources can probably be modeled as separate views of the same information.

Apart from these models we are also planning on integrating a user model, that contains and collects information about the user. This model may be useful in several stages of the system. At a very high level this model can contain information about which services the user likes or disprefers, while at a low level it may contain e.g. information to improve the quality of speech recognition.

5 State of work and current results

I am currently reviewing literature about the state of the art in the area. In parallel to that we are going to perform experiments to simulate interaction with the system and collect data. These experiments are preliminary and serve mainly to develop a robust test setup and explore interaction scenarios and strategies.

We have set up a Wizard-of-Oz scenario that enables us to simulate the surface functions of the system and to record the interaction with users. A schematic overview of this setup is given in figure 3. The room with the dark outline on the right is the test environment. It contains several objects of interest that may serve as subject of discussion with the user as well as a large display.

Participant **C** takes the role of the system user. She is carrying a mobile device that runs the simulation software. The device displays a live video feed showing participant **A** who is playing the role of the animated agent in a separate room. This participant is supplied with a live video feed from a camera on the mobile device that is pointed at the user and picks up facial expressions and gestures. On the wall behind participant **A** a projection displays data or images such as maps. This allows it to simulate the presentation of data by the agent.

Depending on the position of the user, the agent may also be displayed on the display on the left wall in the test environment. A camera is attached to the display that feeds a video stream to the agent participant and allows to direct visual contact between agent and user.

The task of participant **D** is to observe and record the interaction with a mobile camera. This footage is forwarded to participant **B**. This participant simulates the sonification system and produces audio cues depending on the position and actions of the user.

Video and audio streams are saved and will supply us with rich data for evaluation.

6 Future work

We are shortly going to perform first experiments and then begin to evaluate the data, and then use the results to make further design decisions. The experiment setup can later be rened and used to build corpora and collect training data for components that may incorporate machine learning based approaches, such as the dialogue management component.

The LOK8 project is scheduled to run until 2012.

References

[Herzog and Reithinger, 2006] Herzog, G. and Reithinger, N. (2006). The SmartKom architecture: A framework for multimodal dialogue systems. In Wahlster, W., editor, *SmartKom: Foundations of Multimodal Dialogue Systems*, pages 55–70. Springer, Berlin, Heidelberg.

[McTear, 2002] McTear, M. F. (2002). Spoken dialogue technology: enabling the conversational user interface. *ACM Comput. Surv.*, 34(1):90–169.

[Oviatt, 1997] Oviatt, S. (1997). Multimodal interactive maps: designing for human performance. *Hum.-Comput. Interact.*, 12(1):93–129.