A Multimodal Mobile B2B Dialogue Interface on the iPhone

Daniel Porta
German Research Center for AI
Stuhlsatzenhausweg 3
66123 Saarbruecken
Germany
daniel.porta@dfki.de

Daniel Sonntag
German Research Center for AI
Stuhlsatzenhausweg 3
66123 Saarbruecken
Germany
sonntag@dfki.de

Robert Neßelrath
German Research Center for AI
Stuhlsatzenhausweg 3
66123 Saarbruecken
Germany
robert.nesselrath@dfki.de

ABSTRACT
In this paper, we describe a mobile Business-to-Business (B2B) interaction system. The mobile device supports users in accessing a service platform. A multimodal dialogue system allows a business expert to intuitively search and browse for services in a real-world production pipeline. We implemented a distributed client-server dialogue application for natural language speech input and speech output generation. On the mobile device, we implemented a multimodal client application which comprises of a GUI for touch gestures and a three-dimensional visualization. The client is linked to an ontology-based dialogue platform and fully leverages the device's interaction capabilities in order to provide intuitive access to the service platform while on the go.

Categories and Subject Descriptors
H.5.2 [**HCI**]: User Interfaces – Dialogue, Voice I/O, Interaction styles

General Terms
Design, Experimentation, Human Factors

Keywords
Multimodal Interaction, Mobile Business Services, Usability, User Experience, Productivity

1. INTRODUCTION
The success in supporting a company's business processes with a large production pipeline depends on many factors. One of the main factors is the usability of the service you provide. Although mobile business services have great potential to assist individual business experts in general, mobile devices (e.g., mobile phones, PDAs, etc.) are seldom used in important business situations, despite the potential benefits. We should like to point out the fact that implementing a reliable, trustful, and usable mobile interface for business-critical applications is technically challenging and in many cases, the application requirements cannot be met because of device limitations (e.g., limited processing power). We think, however, that when using state-of-the-art mobile device technology, the requirements from the perspective of the business expert can already be met by providing mobile decision support tools.

MobileHCI09, September 15 - 18, 2009, Bonn, Germany.

The approach we explore in this paper makes use of a multimodal mobile client for the iPhone, intended to ease the access to an emerging service platform while on the go. Such platforms, which offer electronic business web services, are becoming more important since we can observe a change from selling monolithic blocks of applications to providing software-as-a-service.

Our mobile business scenario is as follows: Searching on a service platform, an employee of a company has found a suitable service which he needs for only a short period of time for his current work. Since he is not allowed to carry out the purchase, he formally requests the service by writing a ticket in the company-internal Enterprise Resource Planning (ERP) system. In the defined business process, only his superior can approve the request and buy the service. But first, the person in charge has to check for alternative services on the service platform which might be more suitable for the company in terms of quality or cost standards. The person in charge is currently away on business and he carries his mobile device with him that allows him to carry out the transaction while on the go.

This paper discusses the application requirements in more detail (chapter 2) and presents an interaction sequence that we implemented for this B2B scenario on the mobile interaction device (chapter 3). Subsequently, we discuss the technical architecture (chapter 4). Finally, we provide a usability test and a conclusion (chapter 5).

2. REQUIREMENTS
The mobile B2B application is intended to ease the access to an emerging service platform. This platform manages a repository of tradable and composable business services using Semantic Web technologies [1]. It should also support a full service lifecycle from inventing and developing a new service to offering, finding, buying, and using a service. Nowadays, business experts need to stay in touch with the most recent developments to optimize the business processes. Thus, the main requirement is the access to and the interaction with the service platform over the whole lifecycle while on the go.

When relying on standard desktop computers and common web-based interfaces, such a process can be cost and time intensive (the superior is often unavailable). A functional mobile solution, similar to [2], is required to optimize the business processes.

Accessing complex business backend systems with a restricted mobile device poses issues regarding the usability of a mobile client application; it is often not possible to hide the complexity of the former desktop application from the user. On the technical side, we propose a dedicated middleware that invokes the services directly (in contrast to, e.g., [5], where backend services are adapted by providing simplified versions).

In mobile working scenarios, users often have to concentrate on a primary task different from the B2B application; in addition, they might also be in a modality busy setting (for example, in a meeting you cannot use speech input/output) and suffer from an increased cognitive load. An interface that misleads a business expert to make a false decision would hardly be considered useful. Hence, to ensure the safety of the business process, an easy-to-use interface should be provided in the mobile context. [4] describes these requirements as follows: (a) perceived usability of mobile business services, (b) perceived fit for mobile working context, and (c) perceived impacts on mobile work productivity. Accordingly, our mobile business application should reflect the business task but simplify the selection of services and the commitment of a transaction; it should minimize text entry because of the limitations on mobile devices [3]; and it should display relevant information in such a way that a user is able to capture it at first glance.

3. MULTIMODAL INTERACTION SEQUENCE

In order to meet the requirements and to accommodate the limited input and output capabilities of many mobile platforms, we use natural language speech input (including automatic speech recognition (ASR), natural language understanding (NLU)) and output. This omits many tedious steps usually required to search and manipulate (resort and filter) retrieved result lists. Furthermore, speech is appreciated in the mobile context in order to reduce the cognitive load of a user. However, searching for arbitrary, unknown services while using speech (e.g., "Show me services that compute the Eco value of a car seat.") is not supported in our system. For this purpose, we provide a text input field (activated by clicking on the loupe symbol in Figure 3 (c,d)). Instead, speech can be used to refer to the new service (deictic reference). Following our B2B scenario, the user (the superior of the employee) has to decide on the employee's request in terms of a new ticket and launches the client application. The client initially displays an overview of recent tickets.

Figure 1 illustrates the whole multimodal interaction sequence of the mobile user (Figure 1, left) and the dialogue system (Figure 1, right). Technically, they communicate via a central event bus (Figure 1, center) described in more detail in chapter 4. The actual sender of a message transmitted via the event bus to the dialogue system is denoted in parentheses (e.g., GUI, ASR/NLU). The steps are numbered; we refer to these numbers for further explanations. In **(1)**, the user taps on a list entry; in **(2)** the user explores details and in **(3)** and **(4)**, the user explores a result list. Here, joint sort commands like, e.g., "Sort by reliability and price." are also possible. Furthermore, the user can apply a filter, thereby checking one of the company's required quality-of-service standards, e.g., "Which services are faster than 350 ms?". When the user explores the result list, the utterance "And now by rating." implicitly refers to the current discourse context. The missing information is inferred by using a unification technique (Figure 2).

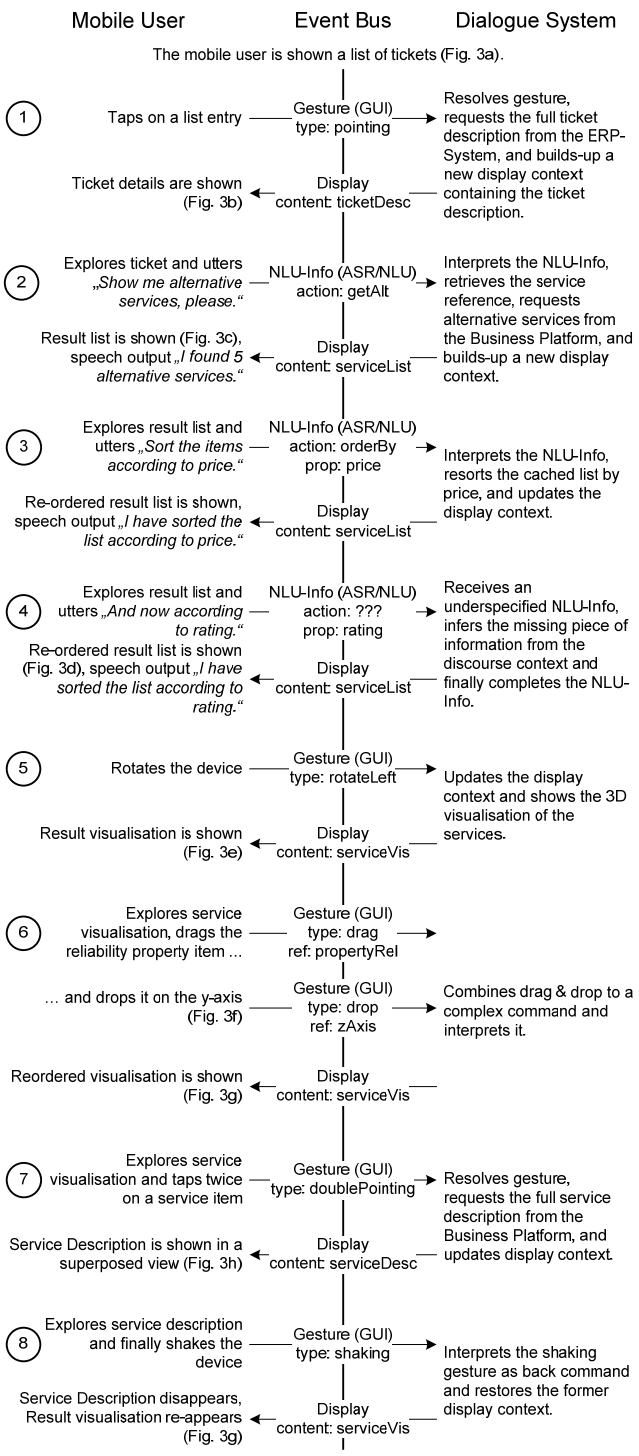

Figure 1: Multimodal interaction sequence

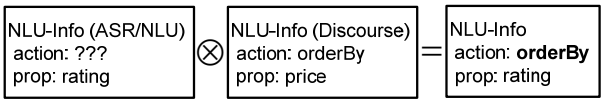

Figure 2: Unification of NLU-Info structures

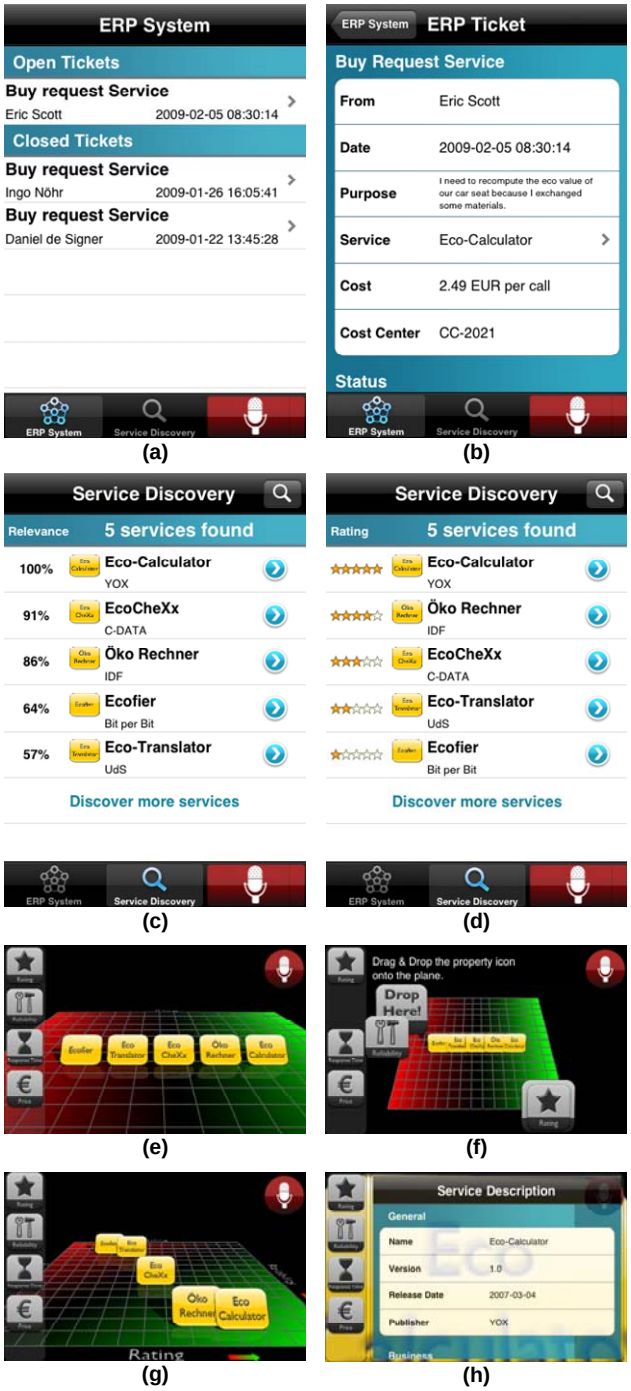

Figure 3: ERP System and Service Discovery Screenshots

In **(5)**, the user rotates the device. The OpenGL-based visualization of the search space arranges the result items in a line, according to the sorting criterion. For zooming and navigating within the environment, the iPhone multitouch gestures can be utilized. We omitted absolute rating values and instead provided the ratio scale with distances and a gradient floor coloring, meaning that "bad" services are positioned on the red part of the floor next to the left corner, whereas "good" services are positioned in the opposite green corner. This ensures that the user finds the best services (with respect to the applied criteria) at the same position. In **(6)**, the user explores the visualization. As an alternative to speech, the user can use drag-and-drop to resort the result items. Previously existing sorting criteria on the x-axis or y-axis are replaced. The service description is shown as a transparent pop-up window **(7)**. It can be discarded by shaking the device (we interpret the built-in accelerometer data) **(8)**. By rotating the device again to portrait orientation, the user can return to the ERP System view and carry out/commit the transaction.

4. TECHNICAL ARCHITECTURE

In order to accommodate the limited processing capabilities of mobile platforms, we use a distributed dialogue system architecture, where every major component can be run on a different server to increase the scalability of the overall system (Figure 4). Thereby, the dialogue system also acts as a middleware between the clients and the business backend services (in order to hide the complexity from the user by presenting aggregated data). There are four major parts, the mobile client, the dialogue system, the business backend, and the event bus, all explained further down.

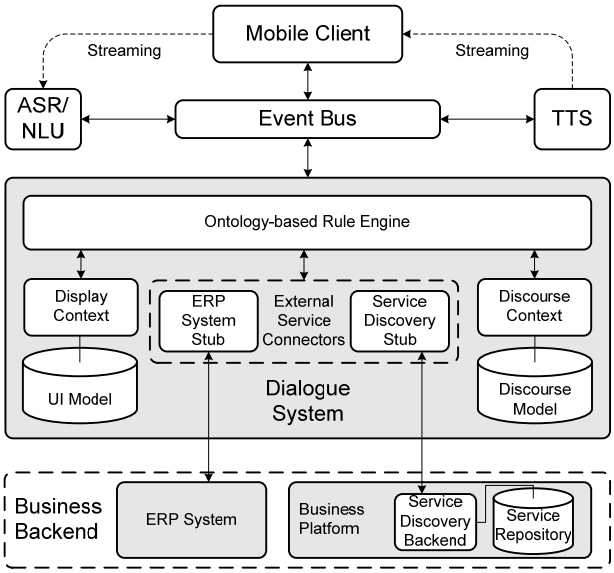

Figure 4: Overall system architecture

4.1.1 Mobile Client

The mobile client is implemented as a native application using the official iPhone SDK. The client provides means to connect to the dialogue system via the event bus, in order to notify it of occurred events, record and playback audio streams, and render the received display data obtained from the dialogue system. In general, the client application is designed as a lightweight component, and the dialogue system is responsible for maintaining the interaction and display context.

4.1.2 Dialogue System

The ontology-based dialogue platform ODP [6] derived from [7] includes interfaces to relevant 3rd-party ASR/NLU (e.g., Nuance) and text-to-speech (TTS, e.g., SVOX) components. It also provides a runtime environment for multimodal dialogue applications supporting advanced dialogical interaction. The central component is a dialogue system which uses a production rule system [8] for a context-aware processing of incoming requests (e.g., display and discourse context) and events. It is based on domain-specific models, e.g., the UI and discourse model. The models include the reaction to pointing gestures, the natural language understanding process, the representation of displayed graphics, and the speech output. Furthermore, the dialogue system provides a programming model for connecting multiple clients (session management) for presentation and interaction purposes. The external and application-specific components in the backend layer can also be accessed easily.

4.1.3 Business Backend

The business backend provides the data to be displayed on the mobile interface. It is accessed by the dialogue system invoking a set of heterogeneous web services by using either the common SOAP or REST protocol. Subsequently, the dialogue system transforms the retrieved data (e.g., XML-based) into its own ontological representation and initiates an update of the current display context.

4.1.4 Event Bus

The main task of the event bus is routing messages between each connected component which currently includes a third-party ASR, a third-party TTS module, and several client applications (i.e., the mobile client and the dialogue system itself). When the multimodal mobile client connects to the event bus, it establishes a new session for the client at the dialogue system. It informs the client about the connection parameters of the ASR and TTS. The speech data is streamed from/to the device in order to ensure fast reaction times. Since we use push-to-activate for the microphone (the user activates the microphone manually), a typical message flow for speech interaction is as follows:

(1) The user pushes the microphone button on the GUI.

(2) The client sends a respective pointing gesture event via the event bus to the dialogue system.

(3) The dialogue system resolves the pointing gesture as “open the microphone” and informs the ASR/NLU via the event bus that it should prepare for speech input.

(4) The ASR/NLU acknowledges this to the dialogue system, which in turn notifies the client that recording and streaming can now begin (on the client GUI, the microphone button turns green).

(5) The user can talk to the client. Upon successful recognition of a spoken phrase, the ASR/NLU sends the recognition result (as NLU-Info structure) to the dialogue system.

(6) The dialogue system informs both the ASR and the client to stop the recording and close the microphone (the microphone button turns red again).

(7) Finally, the dialogue system processes the result.

5. USABILITY TEST AND CONCLUSION

Many previous studies on (mobile) multimodal interaction have focused on a range of aspects such as efficiency gains and recognition accuracy. Only in the most recent studies the user is mobile and subject to the environmental conditions. However, usability can have a quite simple definition in B2B dialogue situations—it means that the (multimodal) mobile interface should enable experts to concentrate on their tasks and do real work, rather than paying too much attention to the interface.

A lot of usability engineering methods are around; we performed user and task observations, and usability testing of our interface where we observed users performing the tasks described in chapter 3. The main objective was to weigh functional correctness higher than the efficiency of the interaction.

We tested this issue in 12 business-related subtasks (similar to the subtasks in figure 1). Eleven participants were recruited from a set of 50 people who responded to our request (only that fraction was found suitable). The selected people were all students (most of them business and economics). Six of them were male, five female and they were partly acquainted with mobile phones. Four of them possess an iPhone/iPod touch. After five minutes of free exploration time with the application, and additional hints from the instructor, users had two attempts to successfully perform a task. For the second attempt, the instructor was allowed to give assistance in terms of clarifying the purpose of the task and the possible input. (If the instructor told exactly what to do, the subtask could not be regarded as performed successfully.)

From our analysis of the questionnaires we conclude that our mobile B2B system can be valuable for the business users. Almost all users did not only successfully complete the subtasks (89% of a total of 132 subtasks), but many of them also provided positive feedback that they felt confident about the ticket purchase being successful. This means the “power test users”, who are knowledgeable of the domain, were able to use the mobile interface for the domain-specific task. We also achieved high query recognition accuracy (> 90%) for the spoken queries that can be recognized (i.e., utterances that are modeled in the speech recognizer grammar).

The current implementation has its limitations, though. Despite the obvious advantage of supporting speech input, flexible language input remains a challenge. All users reported about difficulties when it came to finding appropriate speech commands before they got used to this input modality. In addition, it was rather unclear in which situations a speech command can be used. In the context of the 3-D visualization, eight users reported about problems with the drag-and-drop functionality on the touchscreen. Often, the users (7 of 11) were not able to capture the sense of multiple sorting criteria; many users reported that the icons were just too small. On the contrary, the list representation of the service descriptions was perceived very intuitive while discovering different services. Here, we were able to display a proper description of the background services under investigation. This is due to the fact that every new service is properly described (semantic web services) in a semi-automatic fashion when added to the service repository. In this way, we avoided the problem of displaying search surrogates on the mobile screen [9].

In the second phase of the implementation and testing cycle, we will take the users' feedback into account that we obtained in the first usability test. Subsequently, we will focus on a comparative evaluation of the refined easy-to-use-service in the real business environment to prove that it can be a beneficial and cost-effective solution in addition to pure GUI-based approaches. Our future investigations will also include more fine-grained co-ordination of multimodal input and output.

ACKNOWLEDGMENTS

This research has been supported by the THESEUS Research Programme in the Core Technology Cluster WP4 and the TEXO use case, which was funded by the German Federal Ministry of Economy and Technology under the promotional reference "01MQ07012". The authors take the responsibility for the contents.

6. REFERENCES

[1] Fensel, D., Hendler, J. A., Lieberman, H., and Wahlster, W.: Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential. The MIT Press, (2008).

[2] Sonntag. D., Engel; R., Herzog, G., Pfalzgraf, A., Pfleger, N., Romanelli, M., Reithinger, N.: SmartWeb Handheld - Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services. In: LNAI Special Volume on Human Computing, 4451, Springer, pp. 272-295, (2007).

[3] Kalakota, R. and Robinson, M.: M-Business: the Race to Mobility. 1st. McGraw-Hill Professional, (2001).

[4] Vuolle, M., Tiainen, M., Kallio, T., Vainio, T., Kulju, M., and Wigelius, H.: Developing a questionnaire for measuring mobile business service experience. In: Proceedings of MobileHCI, (2008).

[5] Wu, H., Grégoire, J., Mrass, E., Fung, C., and Haslani, F.: MoTaskit: a personal task-centric tool for service accesses from mobile phones. In: Proceedings of MobMid, Embracing the Personal Communication Device, (2008).

[6] Schehl, J., Pfalzgraf, A., Pfleger, N., and Steigner, J. The babbleTunes system: talk to your ipod!. In Proceedings of the 10th international Conference on Multimodal interfaces. IMCI '08. ACM, New York, NY, (2008).

[7] Reithinger N. and Sonntag D.: An Integration Framework for a Mobile Multimodal Dialogue System Accessing the Semantic Web. In Proceedings of the 9th Eurospeech/Interspeech, Lisbon, Portugal, (2005).

[8] Pfleger, N.: Context based multimodal fusion. In Proceedings of the 6th international Conference on Multimodal interfaces. ICMI '04. ACM, New York, NY, (2004)

[9] Jones, S., Jones, M., and Deo S.: Using keyphrases as search result surrogates on small screen devices. Personal Ubiquitous Comput. 8, 1, pp. 55-68, (2004).