DEVELOPMENT OF A NATURAL LANGUAGE SPEECH DIALOGUE SYSTEM FOR AN AR-BASED, ADAPTIVE MOBILITY AGENT

Ivan Kraljevski, Marion Fischer, Aleksandar Gjoreski, Diane Hirschfeld

voice INTER connect GmbH, Ammonstraße 35, D-01067 Dresden, Germany
{kraljevski, fischer, gjoreski, hirschfeld}@voiceinterconnect.de

Abstract: This paper presents the design and development of a speech interaction module in an adaptive mobility agent based on Augmented Reality (AR). The architecture of the agent combines different components, providing interaction via AR-glasses and speech dialogue, synchronized by a local central management component. One of the core system components is the Speech User Interface (SUI) for hands-free natural language interaction. To provide system adaptability, a User Modelling component as a modality provides additional information about the user's preferences and behaviour, predefined or learned from the recent interactions or former history. Wizard of Oz experiments were carried out in a virtual reality environment providing valuable insights how speakers interact with the system in some predefined scenarios. The experiments also provided knowledge which was used as a basis for the initialisation of the User Model and the final dialogue design, particularly focusing on prompting and recovery strategies.

1 Introduction

Augmented Reality (AR) combines and aligns real and virtual objects with each other in a real environment, and runs interactively in three dimensions and in real time [1]. It has been used in various products and applications for many years, recently mostly in form of applications for tablets and smartphones where the live video feed is overlaid with computer generated graphical elements.

A personal travel assistant or mobility agent which employs AR, overcomes the issues of traditional navigation applications where the routes are presented on 2D maps or perspective view. However, both presentations are non-intuitive and require the user to mentally map street names and symbols on the navigation map to real-world streets and landmarks. AR offers a new perspective through which directions and the distances to landmarks may be displayed in a more intuitive way when the camera is focused on the environment [2], so that the user may simply use actual images of landmarks from the video feed to correlate waypoints' relative position in the real world.

Early implementations of wearable AR navigation systems are presented in [3], where a wearable computer system with a see-through display, digital compass, and a differential GPS are used to provide visual cues while performing a standard orienteering task. While the study [4] describes experimental mobile augmented reality system which employs different user interfaces to allow outdoor and indoor users to access and manage information that is spatially registered with the real world.

For many years, the problem was that AR applications require cumbersome equipment which discourages their use. In order to essentially ease the perception of digital information and to naturally interact with the pervasive computing landscape, the required AR equipment has to be seamlessly integrated into the user's natural environment [5-6]. With the technological advancement and the availability of smart wearable devices as the smartphones, headsets or watches, AR application for navigation and information retrieval matured beyond the stage of a prototype [1] and becomes a new and exciting kind of human computer interface.

An AR mobility agent supports the users in various situations they are facing while travelling to a specific destination. Its tasks include outdoor/indoor routing and navigation, search and information retrieval for routes and points of interest, timetable information and notifications, current traffic conditions, exchanging means of transportation, reporting unexpected changes of the planned route (delays, accidents, traffic jams, cancelled flights, etc.) and many others.

In order to achieve the above mentioned functionality in real time, such system should take advantage of local and remote services and derive higher-order knowledge to notify the user via audio-visual feedback or adapt the current behaviour of the system. Since AR no longer takes place only in two-dimensional space, but in three-dimensional space through the placement of 3D holograms (also called "mixed reality"), development of innovative interaction techniques is necessary. Many of the AR headsets are controlled by integrated basic gesture recognition and if the hands have to remain free for other tasks, then voice control is an adequate alternative.

In this paper, we present the process of design and development of a speech interaction module in an AR-based, adaptive mobility agent, within the RadAR+[1] project. The architecture of the agent combines separate services, including hands-free interaction via AR headset and a speech dialogue.

2 Speech User Interface

One of the core system components is the Speech User Interface (SUI) providing hands-free natural language interaction. SUIs have already found their way into various products. They typically contain Automatic Speech Recognition (ASR) module which estimates set of word hypotheses that best fit a given speech signal according to acoustic and language models. Natural Language Understanding (NLU) module attempts to extract the meaning and the information contained in the uttered phrase. The Dialogue Manager (DM) component controls the dialogue flow with the user and decides what information is presented as a feedback.

Implementing such functionality and quality of interaction on a mobile device is a challenge due to non-permanent connection to remote speech services and limited computational resources. To provide an acceptable user experience, the SUI has to employ robust hybrid ASR with low computational and storage footprint, equally capable of reliably interpreting semantics both locally and remotely.

In the assumed scenarios, the system will be used in outdoor and indoor spaces with a significant presence of background noise, where the user presumably might be under stress and interacting using spontaneous speech (Lombard effect, hyper-articulation, hesitations, breathing, etc.). Therefore, the DM has to be able to handle possible miscommunication by employing adaptive management strategies, using confirmations and error recovery. Also, it has to be able to resolve deictic expressions, pronouns, other anaphora and ellipsis by using other available modalities, like for example the User Model (UM).

2.1 System Architecture

The concept involves real-time exchange of multimodal data provided by many elementary sensors/services, therefore communication protocols as found in Internet of Things (IoT) are suitable for lightweight message transfer. Different modules and services distributed over a range of devices (headset, smart glasses and smartphone) interact by exchanging Message Queuing Telemetry Transport (MQTT) messages over a Hub infrastructure (the MQTT broker). The system architecture is presented on Figure 1.

[1] Reiseassistenzsystem für dynamische Umgebungen auf Basis von Augmented Reality - http://radarplus.de/

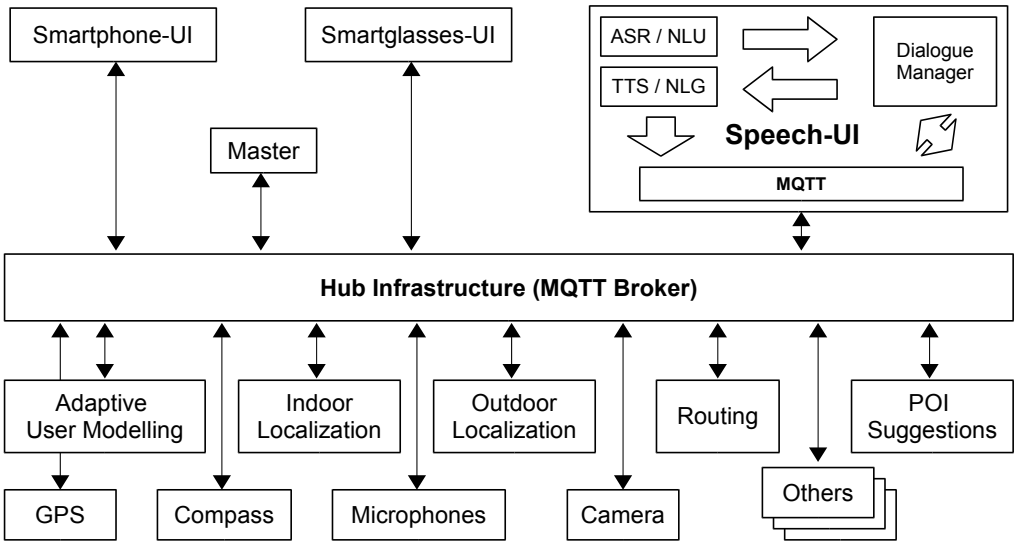


Figure 1. The overall system architecture and the communication between the components

A component sends data with a predefined topic to the MQTT broker which broadcasts it further to the subscribed components. The SUI is handling the incoming messages for the subscribed events. In the case of speech interaction, it will start the dialog to collect information and broadcast back a message with the speaker's intention and corresponding slot values. Based on this modular architecture, the system can be easily adapted and extended by activating/deactivating components which are not necessary or permitted in some particular situations (visual input at airport security checks) or are not available at the moment (no network connectivity).

3 WOz Experiments

In order to collect data and simulate the human-computer interaction in reality, Wizard of Oz (WOz) experiments were carried out in a VR environment, providing valuable insights how speakers interact in some predefined scenarios while covering complete system's functionality. The preliminary investigations using online questionnaires yielded general directions about the preparation of the WOz experiments from the aspect of the users' expectations. Preliminary evaluations [7] with limited SUI functionality involving 12 participants were the basis for the preparation of more complex and elaborate WOz experiments.

The scenarios were based on a predefined list of system's functions described by specific semantic categories like location, time information, mean of transport, points of interest, etc. The experiments also provided a significant amount of multi-modal data, which in the case of the speech interface component is used for creation and optimization of the recognition resources - statistical language models (SLM) and the rule based semantic parser (SEM). The knowledge which was used as a basis for the initialisation of the User Model and the final dialogue design, particularly focusing on prompting and recovery strategies, was also derived from the recorded speaker-system dialogue flow.

3.1 Preparation

To cover all system functions including specific semantic categories, different scenarios were defined. The test scenarios contain tasks the user should fulfil like: discovering the basic functions of the SUI, starting a navigation, asking for help and reacting to system provoked errors and miscommunication. Using a custom WOz tool these scenarios were described as complete dialogues with corresponding prompts to elicit responses from the speaker, confirm his intentions and handle misrecognitions, allowing simulation of a real speech interaction.

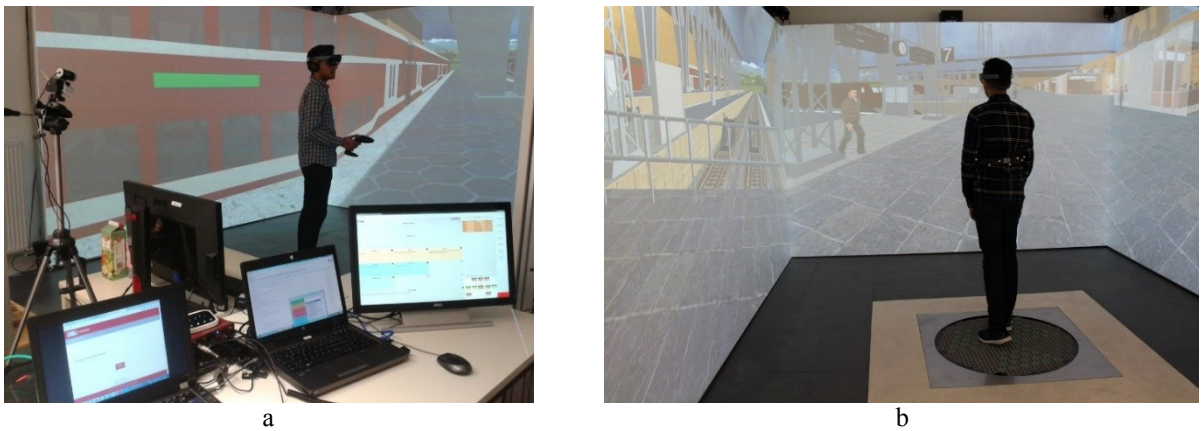
In addition, a questionnaire was created which had to be answered by the participants after completion of the WOz experiments. This questionnaire should pinpoint the possible issues and problems of the speech dialogue system according to the users' expectations in order to improve the human-computer interaction.

Some studies showed, that despite the fact more and more people use speech assistants, the percentage of the voice queries is still rather small (only 20% of queries using Google Voice) [8]. On the other hand, sources reported that 62% used speech interface (commands and searches) at least once in their life, most of them on a smartphone (60%) or navigation device (28%). However, people tend to use voice assistants even less in public and crowded areas, as also reported in [9], mainly due to degradation of speech recognition performance and privacy issues.

To get the advantage of the WOz experiments and to make the later evaluation easier, the sessions and the questionnaire answering stage were audio and video recorded. Simultaneously, the actions of the wizard operating the WOz tool were logged for later dialogue flow analysis. For the WOz experiments, 22 German native male and female speakers between 19 and 67 years of age were recorded.

3.2 Experiments

The WOz experiments took place in a VR environment at the premises of Human-Machine Systems Engineering Department at the University of Kassel. The participants were placed in a room-sized cube called CAVE (Cave Automatic Virtual Environment) where projectors are directed to at least three up to six walls (Figure 2). The projection simulated a virtual train station with rail tracks, trains, people, shops and restaurants, respectively. Sounds and noise of a real train station were fed into the simulated environment to enhance participant's experience. During the WOz experiments, the participants used an AR-headset (Microsoft HoloLens) to get information about timetables, POI-lists and navigation details. The participant had to solve a set of tasks presented by the moderator.



a b

Figure 2. The view of the CAVE room: a. the wizard console, b. participant during a session

The wizard simulated the Speech User Interface by answering the user input with the help of a WOz tool. Miscommunication was simulated by embedding error speech prompts of the following categories: substitution: wrongly recognized parameters; insertion: confirmation of non-uttered sentences and deletion: request to repeat the last sentence.

The error prompts were not triggered automatically, but manually by the wizard. However, not all planned error prompts were played since the actual dialogue flow never reached intended states where the errors should be introduced.

3.3 Transcriptions

The tasks presented in the experiment were aimed to develop Natural Language Understanding (NLU) resources for the speech dialogue, not only regarding speech recognition models but also to improve the audio feedback by carefully designed speech prompts. The speech was segmented and orthographically transcribed by a native speaker, producing detailed descriptions of dialogue flows for each scenario.

Since the participants used spontaneous speech and almost always different sentence constructions or requests to fulfil the same tasks, the transcription process could not be easily automated. Specific recommendations for transcribing were introduced in order to annotate the speaker's intention according the predefined set of tasks described by intents and slots (Table 1). Sometimes, the utterance could be annotated with more than one intent and different slots. The orthographically transcribed recordings along with the semantical annotations were used to improve and optimize the NLU capabilities in order to reflect the real world human-system interactions.

Table 1. Examples of speaker requests and the corresponding transcription from WOz experiments

Spoken utterance	Annotated utterance
Ich möchte bitte nur wichtige Informationen zu meiner Reise sehen.	<intent:user_model_set> Ich möchte bitte nur <value:wichtige> <name:Informationen> zu meiner Reise sehen.
Ich möchte nach Bremen äh nach Kiel Hauptbahnhof über Bremen.	<intent:routing_general_request> Ich möchte nach Bremen äh nach <destination:Kiel Hauptbahnhof> über <via:Bremen>.

The following analysis of the questionnaire found out, that during a navigation 36% of the users want to have speech output only if navigation direction changes against 64% that do not want any speech output at all. In the case of a selection list - like choice of different routes for navigation - 41% of the participants prefer presenting the list via speech and AR output, but 59% prefer just AR output. To select one of the navigation alternatives, 45% of the participants would use the speech input, 41% used gazing and 14% gesture. Therefore SUI implementation in an AR based mobility agent is not a trivial task and should be highly adaptable to the current context and user's preferences or behaviour.

4 NLU Speech Dialogue

It was evident from the results of the WOz tests that the speakers prefer to use a natural way to interact by combining speech with other input modalities. They mostly interacted with the system using unrestricted spontaneous speech. This imposed problems to be addressed, particularly deictic expressions, anaphora, named entity detection, word sense disambiguation, segmentation, etc. These are not easy to resolve, particularly in case of an embedded system with restricted computational resources and non-permanent network connection. In order to provide a high level of NLU interaction, to resolve ambiguity and identify the user's intention, cloud voice services could be employed as a fall-back solution. They provide unparalleled performance in the contrast of domain restricted embedded speech recognition, but in this case they are not always available.

Given the nature of the tasks defined for the mobility agent as, for example, routing and navigation, the system has to be able to identify complex surface expressions and map them to

a particular meaning. For example, non-complete location descriptions in arbitrary order (e.g. only the street name of the current city, without the house numbers, etc.) which could be interpreted as more than one speaker's intention. Such functionality is not possible to achieve by employing handcrafted context-free grammars (CFG), on the other hand a statistical language model of an in-domain text corpus will not cover all the possible sentence constructions (missing flowery phrases, slot synonyms, etc.) without additional effort for adaptation. Additional problem is creating syntax parsers with semantical rules to provide a frame output containing user's intention with its corresponding functional entities (slots).

A common solution is to use word-class language modelling where the classes correspond with the intention slots as defined in the dialogue specification. A word class could be described not only as a list of class elements, but also by a more complex definition, like CFGs or large lists of hierarchically dependent fields (e.g. location addresses, or public transport stations, etc.).

The dialogues are designed using custom (vicDM-Designer) modelling toolkit by defining a list of intents, representing the tasks to be fulfilled by the SUI component of the AR mobility agent, which are compatible with the intents of the other system's components (GUI, AR). Each intent encompasses a number of slots to be filled in order to complete the task. The slots are defined as a list of elements where each has a language specific (in this case German) canonical form and optionally a list of synonyms. Alternatively, a slot could be defined by external pre-build grammars with more complex rules, e.g. for handling numbers, time and date or location (destination, origin) in the format of hierarchically dependent fields (city, address, numbers).

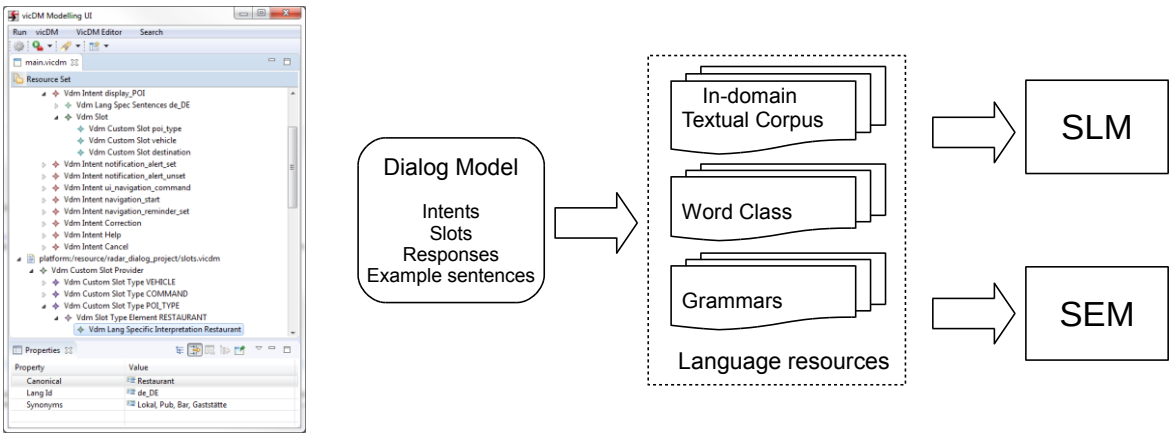


Figure 3. Automatic generation of the speech recognition resources from a dialogue model

The intents are additionally described by example sentences with tagged slots and functional words. These serve as input for automatic generation of the language resources: in-domain textual corpus, word class lists, generated and prebuilt external grammars (Figure 3), which are the basis for building of the SLM and the SEM components (Table 2).

Table 2. Examples of: text corpus (left), ARPA LM (right top) and semantic rules (right bottom)

Ich möchte TIME von ORIGIN zum DESTINATION Ich will nach DESTINATION Ich möchte zu DESTINATION Ich würde gern von ORIGIN nach DESTINATION Aber ich möchte über VIA Aber über VIA Ich möchte nach DESTINATION über VIA	-0.1890562 nach DESTINATION </s> -1.106165 nach DESTINATION über -0.6538671 zu DESTINATION </s> -0.8196934 zum DESTINATION </s>
	<routing_general_request> : (@* <TIME> (von) <ORIGIN> (zum) <DESTINATION> @* | @* (nach) <DESTINATION> @* | @* (von) <ORIGIN> (nach) <DESTINATION> @* | @* (nach) <DESTINATION> (über) <VIA> @*)

With the created SLM+SEM resources the SUI is capable to recognize spontaneous speech and correctly identify the intentions along with the slot values. The performance will be improved by providing a larger set of example sentences and avoiding ambiguities by careful definition of intents.

A prototype system will be used in real world trials to collect additional speech data and to identify out of the vocabulary and out of domain utterances. Those can be identified by a presence of recovery dialogue turns, low confidence recognition results or absence of semantic interpretation. Such utterances are valuable input for continuous expansion and optimization of the dialogue model and he language resources.

5 User Model

In order to achieve system adaptability and to resolve unspecified user input, a User Modelling (UM) component as a modality provides additional information about the user's preferences and habits. They could be predefined or learned both from the recent interactions and from long term history. Adaptation to a user, situation and dynamically changing domains is possible.

The user modelling has been a field of research for a relatively long time [10]. There are many commercial systems deployed years ago profiling their users and giving them recommendations. Online shops are one example where this kind of systems are widely deployed. But not much of the research of user modelling has been applied to speech dialogue systems [11-12].

As the systems that include user modelling usually collect large amounts of data, the data processing is traditionally done on powerful servers. Storing user's information on a remote server leads to privacy issues and discourage the users of using some of the system features [13]. Because in the RadAR+ project user privacy is taken very seriously the whole knowledge about the user is stored locally, on the user's own device. This way, extracting knowledge from the user interactions on isolated embedded or mobile device becomes an important challenge for the system implementation. The vast majority of the current research is not considering these constraints, making the years-old, already proven concepts hard to apply.

When one system with a centralized knowledge database has multiple users, it might be able to extract additional information from the group of users and share the knowledge between them, speeding up the time to learn each user's characteristics. In local embedded or mobile systems there is usually only one user and this kind of help for the knowledge extraction is not available to the system.

To get some insights into the user's habits in an isolated system and avoid cold start, questionnaires before the first system usage are introduced. Data from a questionnaire can be expected to be much more reliable than the learned, implicitly given knowledge. Initialising the User model in this way takes considerable effort on the first start, and might not be acceptable to many users. The learning in this system is designed to be transparent to the user and not to take any unnecessary actions.

The system offers recommendations for route changes, POI to bridge waiting times, and completes underspecified user input by querying earlier user actions and successfully completed transactions. According to earlier user actions and habits the system can make suggestions.

How to present system recommendations via speech interface is a challenge. GUI-based systems are usually giving suggestions slightly out of focus of the main activity, so even if the recommendations are not correct, the user can ignore them and continue with the main activity. In speech dialogues, all information is in the focus. A special care needs to be taken so only the most probable recommendations to be presented, without interrupting the current activity.

Once the recommendations are being presented (implicitly or explicitly), the system is monitoring the user's reaction and see if the given recommendation was accepted. This can additionally strengthen the current user profile, resulting in the user being able to get to the intended action more quickly, with shorter interaction.

6 Conclusions

This paper presents a concept of a SUI module in an adaptive AR mobility agent. The architecture of the agent combines different components synchronized by a local central management component. Wizard of Oz experiments were carried out providing valuable insight how speakers interact in the predefined scenarios and providing the basis for the NLU resources creation and optimization. The analysis results showed that incorporating SUI in an AR based mobility agent is not a trivial task and that the system should be highly user adaptable. To provide system adaptability, a User Modelling component learns from user's actions the preferences and behaviour and can be requested in case of underspecified user input to solve ambiguities and complete a task without annoying question-answering in short time.

Acknowledgement: Research reported in this paper was supported by BMBF (16SV7282)

References

[1] VAN KREVELEN, D. W. F., RONALD POELMAN. *"A survey of augmented reality technologies, applications and limitations." International Journal of Virtual Reality* 9.2 (2010): 1.

[2] SCHUSTER, J.: *"Bahn-App mit Haltestellen-Radar"*, 17.12.2010, http://heise.de/-1155532, 23.07.2015

[3] THOMAS, BRUCE, ET AL. *"A wearable computer system with augmented reality to support terrestrial navigation." Wearable Computers,* 1998. Digest of Papers. Second International Symposium on. IEEE, 1998.

[4] HÖLLERER, TOBIAS, ET AL. *"Exploring MARS: developing indoor and outdoor user interfaces to a mobile augmented reality system".* Computers & Graphics 23.6 (1999): 779-785

[5] NARZT, WOLFGANG, ET AL. "*Augmented reality navigation systems*". Universal Access in the Information Society 4.3 (2006): 177-187.

[6] REHRL, KARL, ET AL. "*Pedestrian navigation with augmented reality, voice and digital map: final results from an in situ field study assessing performance and user experience*". Journal of Location Based Services 8.2 (2014): 75-96.

[7] EIS, ANDREA, ET AL. "*Szenariobasierter Prototyp für ein Reiseassistenzsystem mit Datenbrillen*". Mensch und Computer 2017-Tagungsband (2017).

[8] STERLING, GREG. "*Google says 20 percent of mobile queries are voice searches. Voice search growing as virtual assistant market heats up.*" 18.05.2016, https://searchengineland.com/google-reveals-20-percent-queries-voice-queries-249917, 20.10.2016

[9] KLOSE, ELISA MARIA, ET AL. "*Nutzerorientierte Anforderungsanalyse für ein adaptiv lernendes Reiseassistenzsystem mit Datenbrillen*". Zeitschrift für Arbeitswissenschaft (2017): 1-10.

[10] RICH, ELAINE. *"Users are individuals: individualizing user models"*. International journal of man-machine studies 18.3 (1983): 199-214.

[11] WOLFGANG WAHLSTER (HRSG.). "*SmartKom: Foundations of Multimodal Dialogue Systems. Cognitive Technologies"*, Springer, Berlin, Heidelberg, New York, 2006.

[12] MATTHIAS BEZOLD, WOLFGANG MINKER: *"Adaptive Multimodal Interactive Systems"*. Springer, Berlin, 2011.

[13] SUTANTO, JULIANA, ET AL. "*Addressing the Personalization-Privacy Paradox: An Empirical Assessment from a Field Experiment on Smartphone Users*". Mis Quarterly 37.4 (2013)