Designing, Building and Evaluating Voice User Interfaces for the Home

Stephan Schlögl
Institut Mines-Télécom
Télécom ParisTech
Paris, France
schlogl@telecom-paristech.fr

Pierrick Milhorat
Institut Mines-Télécom
Télécom ParisTech
Paris, France
milhorat@telecom-paristech.fr

Gérard Chollet
Institut Mines-Télécom
Télécom ParisTech
Paris, France
chollet@telecom-paristech.fr

CHI'13, April 27 – May 2, 2013, Paris, France.
ACM 978-1-XXXX-XXXX-X/XX/XX.

Abstract
Speech-based human-device interaction is booming and companies such as Apple, Google and Nuance constantly provide us with potentially new application scenarios where voice may be the control modality of choice. Appropriate methods for designing and evaluating these voice-operated applications are, however, rare. This paper reports on our attempt of combining different Language Technology Components with a web-based Wizard of Oz tool to design and prototype a list of home care and communication services. Guided by a research project that aims at providing applications optimized for seniors, we describe our planned approach of combining wizard actions with technology output in order to build and evaluate applications that use natural language as an interaction paradigm.

Author Keywords
Prototyping; Voice User Interfaces; Wizard of Oz

ACM Classification Keywords
H.5.2 [User Interfaces]: Natural language; H.5.2 [User Interfaces]: Prototyping; D.5.2 [User Interfaces]: Voice I/O

General Terms
Design, Human Factors

Introduction

Products such as Apple's *Siri*[1] and Google's *Voice Search*[2] demonstrate how Voice User Interfaces (VUI) have the potential of changing the way we interact with our computing devices. While for some talking to a computer may simply convey a great user experience, for others, it can offer a significant alleviation when interacting with a piece of technology. In 2010 17.38% of Europe's population was older than 65 years of age and current projections suggest that by 2060 we will have less than two people of working age (15-65 years) for every person beyond 65 [2]. Technologies that offer more natural and less cognitive demanding interaction channels, such as speech, may therefore not only attract our technophile young but generally be seen as a way of supporting the life of an aging society. Yet, building, integrating and combining language technologies such as speech recognition, natural language understanding and speech synthesis, for different application scenarios, poses significant design and software engineering challenges.

In this paper we present work that is conducted as part of the vAssist[3] project. The goal of vAssist is to support seniors suffering from (fine) motor skills impairments and/or chronic diseases in their everyday life by offering voice-based interaction capabilities for a list of home care and communication services. These VUIs are implemented following a Human-Centered Design approach where end-users (and their living environment) are involved in all stages of development, assuring that features and interaction paradigms are adapted to people's requirements and specific needs.

[1] http://www.apple.com/ios/siri/
[2] http://www.google.com/mobile/voice-search/
[3] http://vassist.cure.at/home/

The Challenge of Building VUIs

Unlike the design and implementation of graphical user interfaces (GUI), where the designer (in close connection with the end-user) defines the control language that operates an application, the construction of VUIs mainly depends on domain-specific natural interaction data (i.e. variances of spoken input and dialog behavior). That is, while a button-click on a GUI can be interpreted as a clearly defined signal to initiate an associated action, spoken commands are often highly ambiguous. Even in cases where the interaction is strictly driven by the system, and therefore the user is actively prompted for specific inputs, the actual response may vary significantly. For example, if a calendar application asks for the ending of an appointment, the user might respond with an exact time (e.g. "at 12pm"), with an approximation (e.g. "around noon"), or maybe provide a duration (e.g. "it will last for 2 hours"). It is the voice interface designer who needs to define a set of possible responses so that an appropriate language model can be integrated; a task which can be difficult in particular if the application is situated in a novel, unexplored domain. Building a compelling VUI therefore not only requires capable hardware and efficient machine learning algorithms, but also demands a certain amount of speech and dialog data that is necessary to sufficiently define the potential interaction space. To support this gathering of realistic dialog interactions our lab is currently working on a flexible prototyping framework which combines several open source Language Technology Components (LTC) with a web-based Wizard of Oz experimentation platform. We believe that such a set-up, in which a human wizard is able to simulate as well as augment technology output, should allow for both the collection of linguistic data and the design of voice-based interactions.

A Framework for VUI Design

The processing of spoken natural language usually requires a set of technology components to be aligned in a progressive processing chain. It starts with an Automatic Speech Recognition (ASR) module which converts spoken input into text. Next this text is interpreted by a Natural Language Understanding (NLU) component which extracts the relevant meaning (a task that can be difficult at times). The output of the NLU is then sent to the Dialog Manager (DM) which initiates a suitable response. This response command is forwarded to a Natural Language Generation (NLG) component that produces the relevant text utterance and sends it to the Text-to-Speech Synthesis (TTS) module which finally converts the text into spoken output. As a whole such a set-up is called a Spoken Dialog System (SDS).

Researchers all around the world work on these types of systems and produce components that are often open-source and freely available (at least for research purposes). To build a prototyping framework for (spoken) language-based interaction we have taken the Julius[4] ASR, the Disco DM and the OpenMary TTS[5] and combined them to a (multi-lingual) SDS. In terms of NLU we plan to use our own component which is based on Jurčíček et al.'s natural language interpreter [4]. This component, however, requires domain-specific interaction data to be operational. Thus, in order to collect the necessary data we have further integrated the WebWOZ experimentation platform[6]. As the entire framework is accessible via the Internet, relevant experiments can be conducted in our lab as well as in users' homes (given the availability of broadband Internet).

[4] http://julius.sourceforge.jp/en_index.php/
[5] http://mary.dfki.de/
[6] http://www.webwoz.com/

Wizard of Oz Experimentation

Wizard of OZ (WOZ) constitutes a prototyping method that is used by researchers and designers to obtain early feedback on applications/features that usually require significant resources to be implemented. Speech- and natural language-operated software can be seen as one representative of this type of applications. In a so called 'WOZ experiment' a human 'wizard' simulates the functions of a potential future system, either entirely or in part, which allows for the evaluation of user experiences and interaction strategies without the need for building a fully functional product first [5].

The application area of WOZ prototyping ranges from very low-fidelity evaluations, where researchers might refer to paper prototyping as being a form of WOZ, to high-scaled simulations in which the wizard only takes over a very distinct functionality of an envisioned system. Similarly, the types of experiments may vary. On the one hand we find set-ups where a test participant is openly aware of the simulation, sometimes even taking over the task of the wizard in order to better convey a constructive argument, on the other hand it is often of high importance that participants believe that they are interacting with a real system, so that their actions and consequently their feedback is just as real.

One application area of WOZ, which is of particular interest to us, is its use as an instrument for collecting various types of corpus data. From a VUI development perspective this collection of data may be compared to the production of a GUI design concept and therefore constitutes an essential aspect of a human-centered implementation methodology.

Using WOZ for VUI Design

Following the motivations outlined above we plan to use WOZ for collecting an initial corpus of spoken language. Furthermore we would like to explore dialog strategies as well as the overall user experience of 'talking' to a device/system. As for linguistic corpus gathering our focus lies on producing reliable quantitative data, based on which valid algorithms may be derived. The WOZ experiments should therefore be highly controlled and the wizard be trained to the point where he/she acts consistently. The WOZ tool, which we have slightly adapted to fit our use case scenarios, should help achieving this consistency and also control for confounding factors such as variability in utterances sent to a participant or irregular response timing.

From a participant's perspective, literature has shown that people interact with 'intelligent' systems differently than they do with human interlocutors [3], even if their linguistic capabilities were equal (as in theory it is the case in a WOZ setting). They adapt their language similar to when talking to a child or to a foreigner, they use a smaller set of words, and they are more inclined to wait for responses [1]. While we are aware of these aspects, and will take them into account when interpreting our results, we are also interested in whether a familiar environment, such as the home, and a certain habituation to the application/system can lead to a more natural form of interaction.

Prototyping VUIs for the Home

In order to design VUIs for a list of home care and communication services we are currently preparing a set of user studies to be run in our lab (home-like environment) as well as in users' homes. Participants will be interacting with devices they are already familiar with (e.g. PCs, TVs, mobile phones). Some of the applications we are testing may, however, be new to them (e.g. a well-being diary). Convincing simulations are therefore crucial so that realistic feedback as well as robust linguistic data can be collected. First trial runs of an 'open' WOZ set-up, where test participants were aware of the simulation, have already been conducted and shown that the approach is feasible. However, for the collection of useful linguistic data participants have to believe that they are interacting with a real product. To achieve this we believe that our prototyping framework is the right instrument. The results of an initial set of trial runs should be available by the time this workshop will be taking place and so we would like to take this opportunity and share our impressions as well as first lessons learned with this selected sub-group of the CHI community.

References

[1] Dahlbäck, N., Jöonsson, A., and Ahrenberg, L. Wizard of oz studies - why and how. In *Proceedings of IUI* (1993), 193–200.

[2] European Union. Active ageing and solidarity between generations. In *Eurostat, Statistical books* (2011).

[3] Jönsson, A., and Dahlbäck, N. Talking to a computer is not like talking to your best friend. In *Proceedings of SCAI* (1988), 297–307.

[4] Jurčíček, F., Thomson, B., and Young, S. Reinforcement learning for parameter estimation in statistical spoken dialogue systems. *Computer Speech & Language 26*, 3 (2011), 168–192.

[5] Kelley, J. F. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of CHI* (1983), 193–196.