Speech Interface Dialog with Smart Glasses

Aryan Firouzian and Petri Pulli
Dept. of Information Processing Science,
Faculty of Information Technology
and Electrical Engineering,
University of Oulu,
Oulu, Finland.
Email: Aryan.Firouzian@oulu.fi, Petri.Pulli@oulu.fi

Matus Pleva, Jozef Juhar, and Stanislav Ondas
Dept. of Electronics and Multimedia Communications,
Faculty of Electrical Engineering and Informatics,
Technical university of Kosice,
Letna 9, 04120 Košice, Slovakia.
Email: Matus.Pleva@tuke.sk,
Jozef.Juhar@tuke.sk, Stanislav.Ondas@tuke.sk

***Abstract*—This paper describes design of elderly-user-friendly multi-mode user interface with different modules. Use of eye glasses is common among senior citizens, and it inspired us to implement interface modules on it. Indicator-based Glasses contains the Eye Blinking Detection module integrated with visual cues indicators as system feedback. The multi-mode interface provides five interaction channels by proposing audio input/output modules and Android application on smartphone device. The VoiceXML Dialog Manager implementation (VoiceON) is described and proposed for speech enabled computer initiated dialogues. Senior citizens suffering from mild and moderate dementia are the primary target group of the proposed system. The human factors of the multi-mode interface will be tested in experiment with senior citizens, and different scenarios will be evaluated.**

***Keywords*—Indicator-based Glasses, Automatic Speech Recognition, Dialog Management**

I. Introduction

Kortum [1] studied nontraditional user interaction techniques, and discussed usability improvement in auditory interface, voice dialogue, audio display, interactive voice response (IVR) when they are integrated as a module in multi-interface system. Considering this motivation, Kortum [1] stated that there are two types of multi-interface interaction systems. Multi-mode interface includes two or more interaction techniques, which are designed to perform and accomplish same unique task. The user freely can select the preferred method of interaction technique based on the context. On the other hand, multi-modal interface includes two or more interaction techniques, which are combined to perform and accomplish only one specific task. So, the user has to perform different interaction methods to perform the task. In most cases, the multi-modal interface is designed to resolve the ambiguity problems in the interaction process. [1]

Multi-mode interactions are significantly common among the service providers, when they allow users to access and modify data by the preferred interaction method such mobile user interface, voice interactive response or web browsing. Zhang et al. [2] designed a gaze and speech multi-modal interface to select objects with different colors in environment. Speech and gaze recognizer report accuracy scores, and integrated module resolves inaccuracy problem which may caused by gaze deviation or noisy environment. [2]

Although nontraditional, multi-mode and multi-modal interfaces are different from traditional graphical user interface, they must conform to basic interaction principals such as ISO 9241:11 to provide effective and efficient interfaces to satisfy user expectation. Early testing and experiment with target user group measure usability factors such as effectiveness, efficiency, satisfaction level and error tolerance. [1]

In this study, we explain design of context-sensitive multi-mode user interface with more than one interaction techniques. Touch screen input and output, voice interactive response and blinking detection system are the interaction modules in the proposed system.

Significant achievement in the development of voice-based interfaces through automatic speech recognition (ASR), natural language process and text to speech (TTS) in the recent years open the door to develop more intuitive IVR system. Open source and commercial products are currently available to be used as above-mentioned modules of context-aware IVR system. Cloud-based commercial apps are described in [3].

Prylipko et al. [4] explained the architecture of Zanzibar OpenIVR which includes speech application server utilized with Voice-XML interpreter (JVoiceXML), speech recognition engine (CMU Sphinx 4) and text to speech engine (FreeTTS). The prototype used dialogue management with mixed imitative dialogue strategy. The mixed strategy allows the system to ask for missing information from the user, while still following the form items in VoiceXML. Prylipko et al. [4] also tabulated different characteristics of available platforms for spoken dialogue system. The comparison focuses on use of modular component, VoiceXML dialogue and being open source. [4]

Cohen et al. [5] provided guideline to design Voice User Interaction (VUI) by including prompts, system message, grammars and dialog logic. The dialog logic provides the system action procedure in response to user's actions. Furthermore, Schnelle [6] involved context information to support workflow engine while designing audio-based interface. The context information is collected from the environment such as location and current task in hand. Since the prototype was developed on 2007, limitation on mobile and wearable devices forced the designer to consider a separate desktop server to integrate the tasks for workflow engine, VoiceXML interpreter, ASR, TTS and dialogue manager; while the auditory messages

and commands were streamed to/from end user client. [5, 6]

Although context-aware VUI can be developed to provide human-like conversation and mature man-machine interaction to obtain information from user, in most cases the nature of the tasks defines suitable interaction technique especially if the user suffers from memory decline or physical disability. In some cases, mobile graphical interface can catalyze the interaction while IVR, VUI and auditory display fail to fulfill user expectation. In fact, user perception of human-to-human conversation lead to the high user expectation of VUI system, while the system fail to properly proceed with unpredictable user input, and it is considered as one of the major drawback in VUI systems. Sorri et al. [7] conducted experiment with elderly users with dementia to evaluate usability performance of three different type of user interaction technique as output channel. Auditory display, visual-based information and tactile messages were compared each other to get navigation instruction while performing a simple navigation task. Considering nature of the navigation task, the result indicate that the mixed auditory display and visual cue provide the most efficient interaction to accomplish the tasks. [1, 7]

Our previous study includes design of visual cue navigation instruction system for elderly users and bike riders. The interface provide navigation cues and voice commands through led-based glasses. Figure 1 shows the prototype and example visual cue commands, which can be adjusted based on ambient light, and user preferences. It is suitable for bikers since their hands are not free to interact with hand-held devices. We conducted experiment with young bikers and elderly users to evaluate usability of the prototype in navigation tasks. In case of elderly users, most of the participant suffered from mild or mediate dementia, but they accomplished their navigation task by assistance of visual cues and voice command.

Two unresolved issues are revealed during the interview sessions. First, users expect the glasses to help them with some other tasks in daily activities such as shopping, calling and organizing calendar. Second, they expect the system to preserve their dignity and privacy in public places by providing unobtrusive interaction. Loud voice interaction through VUI is an example that violates the privacy of the users. Furthermore, the speech-enabled system which is designed for elderly users should be customized to be used in public places. [8]

Above-mentioned argument accentuates simple problem of speech/voice dialogue system, which is preserving dignity and privacy of the users in public environment. This is the primary motivation behind this study and it derives us to design a conceptual context-aware multi-mode interaction system that resolves the problem, and bring efficiency and performance to the system together with dignity and privacy to the user.

Eye blinking detection is considered as an unobtrusive input channel to fetch information from users. Many research prototypes and techniques are designed to detect blinking eye action in the recent years. In most cases, a camera with image processing solution is setup such as Pupil Lab eye tracker. Infrared proximity sensor is also popular implementation, which is used in Google Glasses. Biopotential measurement

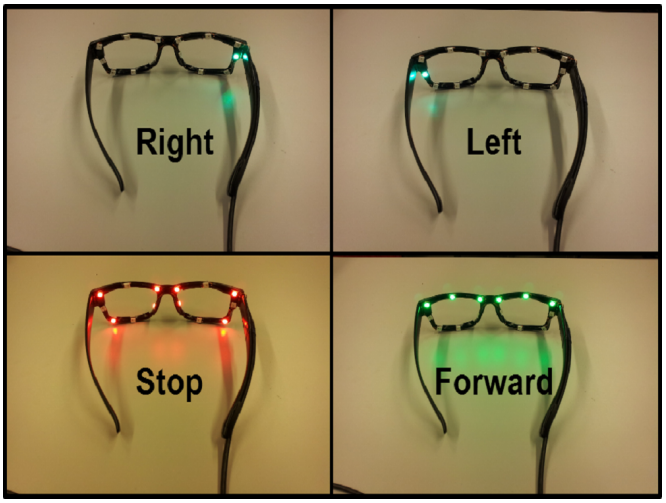

Fig. 1. Indicator based glasses is designed to provide near eye visual cue to assist users in navigation tasks. Android application set brightness, color, blinking frequency and combination of LED indicators.

such as electrooculography (EOG) sensor implementations or EEG [9] are considered as unobtrusive implementation of eye blink detection in eye glasses. Since use of eye glasses is common among the senior citizens, implanting eye blink detection system in eye glasses is the most suitable approach to design elderly-user-friendly interaction. [10, 11, 12, 13, 14, 15]

Zajicek [16] studied speech enable user interface for senior citizens which is known as VoiceXML-based Voice Activated Booking System (VABS). The user interface designed to fetch information from the web regarding bus timetable, council collection, doctor appointments. Furthermore, Zajicek [16] argued that the user interface should follow three design principles, which are:

1) The output message should be short but understandable.
2) User should be able to reduce number of choices as desired.
3) Confirmatory message should be considered in different steps of dialogue communication.

The study concludes with six specific pattern forms of messages, which are menu choice message, confirmatory message, default input message, context sensitive help message, talk through message, explanation message, partitioned input message and error recovery loop. [16]

Elderly user group is the primary subjects in web browsing experiment with VABS. Considering the result of the experiment, our proposed prototype inherits pattern forms classification from characteristics of VABS. Another reason to apply design principals in our multi-mode user interface is the similarity of tasks and possible scenario that our prototype should be able to handle. The user interface should handle the daily tasks of senior citizens efficiently and provide smooth transition between modules. The tasks includes reading and modifying shopping list, appointments and calendar events; reading news; providing navigation instruction and managing the phone calls and text messages. Furthermore, nature of

dialogue-based interface facilitates answering short health, cognitive or nutrition state examination. [16]

II. Architecture

Common use of eye glasses among the senior citizens has derived us to implant interaction modules on wearable glasses. Furthermore, it supports our research motivation which is designing unobtrusive interaction mechanism in public environment. Android application running on a hand-held device works as local processor, and it communicates with headset Bluetooth (voice user interface) and Bluetooth transmitter (eye blink detection). Android application provides graphical user interface and synchronize dialogue sequence for eye blink detection and voice user interface modules. The local processor parses dialogue and communicates with remote processor. Sensory ambient data such as GPS location data are sent to remote processor, and then relevant dialogue is generated and sent back to local processor to initiate the interaction with user. Abstract architecture of the system to provide and parse dialogue document is shown in Figure 2.

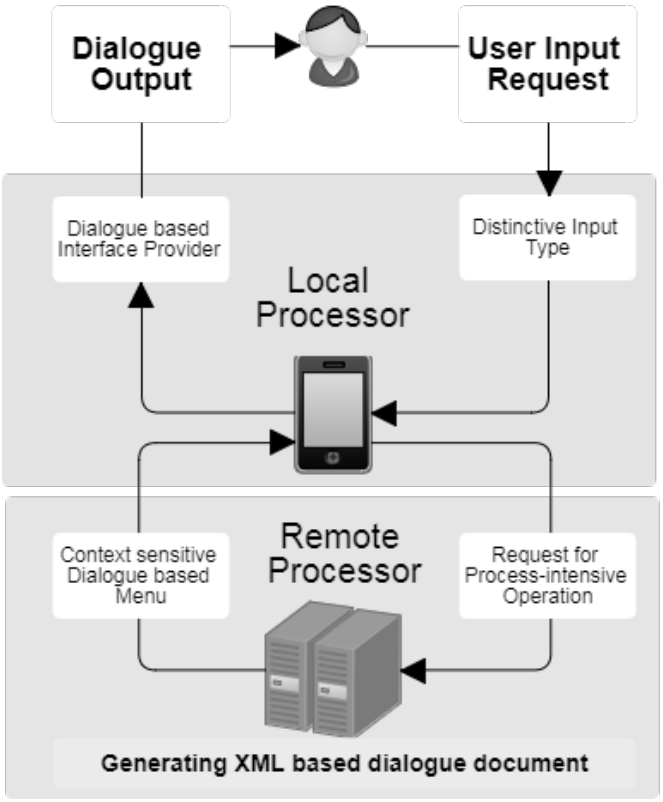

Fig. 2. Abstract architecture of the system.

Voice user interface allows user to interact in convenient approach while hands are involved in other tasks. The speaker and microphone components of Bluetooth headset device are implanted on the side arms of glasses frame to compose hardware input and output channel of voice user interface. Android application activates the interface and speech dialogues are streamed between two endpoints. However, to preserve dignity and privacy of the user in public places , the interaction can be extended to blink detection module for input commands.

The proximity QRD1114 sensor, integrated with infrared LED and phototransistor, is the main hardware component of eye blink detection module. It detects reflected infrared signal from blinking eye and informs Android application through low energy Bluetooth transmitter. The proximity sensor is implanted on the corner lens frame of the glasses to avoid blocking user's field of vision. A rechargeable Li-ion battery supplies the electricity power for Bluetooth transmitter, proximity infrared sensor and Bluetooth headset.

The above-mentioned system architecture provides hardware platform for multiple input/output channels between user and system (3). This study argues the process of generating the dialogues documents, and different types of dialogues (message/command) that can be handled by three modules through five channels.

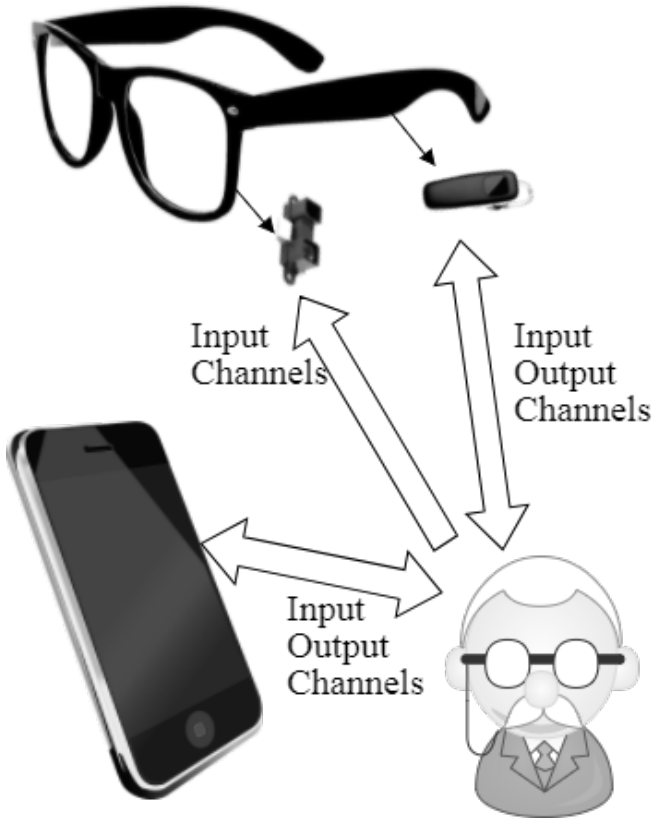

Fig. 3. There are five channels of communication between users and multimode interface: Input and output channels for mobile graphical interface; input and output channels for speech enabled interface; input channel for eye blink detection.

The interaction between user and system is triggered by user request message which presented by voice user and graphical interface. The user request command is sent to the remote server to generate required dialogue. This process is performed through partitioning the dialogue and providing menu choice message.The remote server includes a machine learning engine

to consider user request command, time, location and ambient sensory data as features to learn user behaviour to generate the Voice-XML document and initiate the dialogue. For example, based on user request commands, the dialogue can be generated to manage the shopping list or set an appointment in the calendar.

Voice dialogue management is performed by VoiceON dialogue manager [17], which interprets VoiceXML language. Using of VoiceXML-based solution is well suitable for intended application, because of well defined dialogue flow. Moreover, VoiceXML language has proper readability (see Fig.6), which supports fast development of new services and also automatic generation of VoiceXML code by machine learning algorithms. The architecture of the manager is shown in Fig.4. VoiceOn consists of two main parts - Dialogue manager server and the wrapper module (see Fig.5). This arrangement enables extensive use of dialogue manager for various dialogue systems, because only wrapper module need to be changed. Wrapper translates messages between VoiceON manager and a platform, which contains other modules of communication system as are automatic speech recognition, text-to-speech synthesis [18] and others. The VoiceON system offers two wrappers - one for DARPA Galaxy architecture [19] and one for Aldebaran NAO robot [20].

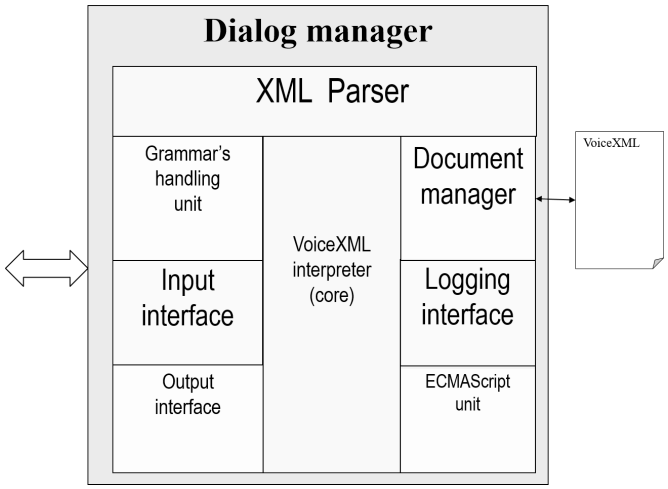

Fig. 4. TUKE VoiceON Dialog manager architecture

VoiceON dialogue manager is written in C++ and supports also subset of elements from W3C SRGS and SISR recommendations, which enable to write complex XML speech grammars with semantic interpretation tags.

The next figure (Fig.5) shows the architecture of designed voice dialogue system, where the glasses are used as input-output interface device. Behind this front-end, voice dialogue system consists of three basic modules - automatic speech recognition (ASR), Text-to-Speech (TTS) and VoiceON dialogue manager.

In order to allow users to select method of interactions based on the context, different type of input/output and types of messages and commands should be defined. Menu choice

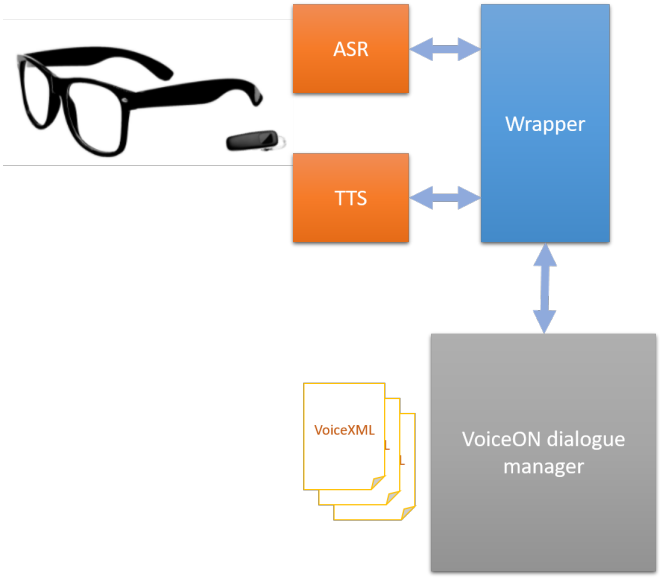

Fig. 5. Architecture of the voice dialog system

```
<?xml version="1.0" encoding="UTF-8"?>
<vxml version="1.0" lang="English" application="applications/ARD.vxml">
<var name="Name"/>
<form id="steptwo">
 <field name="YourName">
<nomatch>
      <prompt bargein="false">Sorry, I don't have your name in my database. </prompt>
 <goto next="#help"/>
   </nomatch>
  <grammar src="lang_resources/grammars/names_eng.xml"
type="application/grammar+xml"/>
 <prompt bargein="false"> Hi. I am your assistant. What is your name? </prompt>
  <filled>
<prompt bargein="false"> Hello <value expr="YourName"/>. </prompt>
<assign name="Name" expr="YourName"/>
<goto next="#help"/>
 </filled>
</field>
</form>
 <form id="help">
  <noinput>
      <prompt> Could you repeat your answer? </prompt>
      <reprompt/>
  </noinput>
```

Fig. 6. VoiceXML code example

dialogue provides several items for selection. There are many studies arguing number of possible items to remember in voice dialogue system. Based on capability of user's short term and working memory, prior study suggest 6 to 7 items for young users and up to 3 items for senior citizens, while the first and last items have higher possibility to be remembered. Providing less items means more interactions, navigation and depth are needed to reach the final goal. However, previous studies consider short messages approach with more in-depth navigation as more efficient approach with elderly users. [16]

Confirmatory dialogue is used to assure user's selection or input interaction. All three input channels can handle confirmatory interaction. It provides user with only two alternative of confirm and rejection.

Default input message is considered as subset of menu

choice message, and it is generated by machine learning engine based on user's behaviour and sensory data to include most fitted item in the menu choice message. If the user fails to interact with the system, the default input message is selected by the system for proceeding to the next task. The main drawback of default input message is the threat of unsatisfactory message which needs frequent users interaction over the time for training the machine algorithm.

Explanation message is provided by specific command and it presents user with extra information related to current task and dialogue. It is triggered by an arbitrary speech command or button in the graphical interface. Explanation message is useful when the system fails to provide an intuitive approach of interaction. Zajicek [16] discussed necessity of context sensitive help message which is a human to human communication, and talk through message which is continuous provision of task instruction to the users. However, we consider them as subset of explanation message, as human assistance can be activated as a extension of explanation message and need for continuous task instruction is highly dependant on the performance of working memory.

Recovering from speech input error required strategy to avoid user frustration, and still focus to the task accomplishment. Partitioned input message is designed to systematically categorized possible inputs from user, similar to a decision tree model. It is triggered by providing menu choice and confirmatory messages in the error recovery loop when the system fails to recognize input speech.

Designing the speech enabled system for elderly facilitates interaction of the user while the system recognizes user input from a set of valid expressions. On the other hand, it brings high risk of usability issue if speech recognition fails. Error recovery process keeps the user on the right track by partitioning valid statements and providing menu choice and confirmatory messages. Using partitioned input message results the trade-off which increase the length of the interaction by providing more dialogues, while it reduce the risk of errors.

User request message, menu choice message, confirmatory message and explanation message are presented to user by audio display and graphical interface in parallel. On the other hand, menu choice command and confirmatory commands can be given by eye blink detection module in addition to other two modules. The Table.I demonstrates modules that are involved in different type of dialogue forms.

TABLE I
MODULES TO PRESENT MESSAGES AND CAPTURE USER'S COMMANDS.

	eye blink detection	voice user interface	graphical interface
user request message		X	X
user request command		X	X
menu choice message		X	X
menu choice command	X	X	X
confirmatory message		X	X
confirmatory command	X	X	X
explanation message		X	X

III. CONCLUSION AND FUTURE WORK

The proposed interaction mechanism is designed to bring two advantages of fault tolerance and unobtrusiveness to the existing speech enabled interaction, and it provides an elderly-user-friendly system.

First, it outlines the procedure of handling speech recognition error, and provision of partitioned message to the user.

Second, it enables context sensitive and convenient interaction which not only customizes the dialogue content, but also provides users with multiple parallel interaction modules to accomplish same task based on the context.

Using dialogue based multi-mode interaction system brings trade off between task complexity and above-mentioned advantages. Furthermore, the main target group of the study is elderly users suffering from mild and moderate dementia, and the system is not expected to handle complex tasks. The proposed solution should be able handle simple tasks such as navigation or managing appointments, shopping list and calls. Furthermore, it provides suitable platform to conduct short examination regarding health, cognitive or nutrition state.

There are several issues relating to internal validity threats in this study. Different technology aspects are involved in the research which should be addressed during the improvement phase. Generating well-partitioned dialogue document needs assessing iterative experimentation and implementing customized practice of design. Inclusion of emotion detection module [21] could improve the system user friendliness.

VoiceXML documents are traditionally used in Dual Tone Multi Frequency (DTMF) to provide telephony services for customers. Since DTMF system are designed for regular users, time interval between choices in the menu forms are not considered as usability concern. Error recovery loop provides strategy to make transition between message types. However, repeating number of message should also be investigated before transition in processing user request command, menu choice command and confirmatory command. Time interval and repeating number of messages are metrics which should be studied in the experimentation phase.

Experimentation phase should be conducted in three blocks to evaluate usability of the system with combination modules of blink detection system, speech enabled module and graphical interface in hand-held mobile device.

Prior studies argued that context related questions must be short and expect short answers. Qualitative studies should be conducted to measure satisfaction level of elderly users with different type of messages considering length and clarity of the messages.

The experiment should be conducted in three different phase with primary subject group. In the first phase, users need to accomplish the task with graphical interface only.

The second phase allows subjects to use speech messages and commands.

In the last phase, users have alternative to interact with all modules of Eye Blink Detection, Voice User Interface and Graphical User Interface based on preference.

ACKNOWLEDGMENT

The research presented in this paper was supported by Academy of Finland and Japan Science and Technology Agency (JST) in ASTS (Assisted Living for Senior Citizens) project and by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the projects VEGA 1/0075/15 & KEGA 055TUKE-4/2016.

REFERENCES

[1] P. Kortum, "HCI beyond the GUI," *Morgan Kauffman, San Francisco*, 2008.

[2] Q. Zhang, A. Imamiya, X. Mao, and K. Go, "A gaze and speech multimodal interface," in *Distributed Computing Systems Workshops, 2004. Proceedings. 24th International Conference on*. IEEE, 2004, pp. 208–213.

[3] J. Collinaszy, M. Bundzel, and I. Zolotova, "Implementation of intelligent software using IBM Watson and Bluemix," *Acta Electrotechnica et Informatica*, vol. 17, no. 1, pp. 58–63, 2017.

[4] D. Prylipko, D. Schnelle-Walka, S. Lord, and A. Wendemuth, "Zanzibar OpenIVR: an open-source framework for development of spoken dialog systems," in *Text, Speech and Dialogue*. Springer Berlin/Heidelberg, 2011, pp. 372–379.

[5] M. H. Cohen, M. H. Cohen, J. P. Giangola, and J. Balogh, *Voice user interface design*. Addison-Wesley Professional, 2004.

[6] D. Schnelle, "Context aware voice user interfaces for workflow support," Ph.D. dissertation, Technische Universität Darmstadt, 2007.

[7] L. Sorri, E. Leinonen, and M. Ervasti, "Wayfinding aid for the elderly with memory disturbances." in *ECIS*, 2011.

[8] A. Firouzian, Y. Kashimoto, Z. Asghar, N. Keranen, G. Yamamoto, and P. Pulli, "Twinkle megane: Near-eye led indicators on glasses for simple and smart navigation in daily life," in *eHealth 360*. Springer, 2017, pp. 17–22.

[9] L. Vokorokos, B. Mados, N. Ádám, and A. Baláz, "Data acquisition in non-invasive brain-computer interface using emotiv epoc neuroheadset," *Acta Electrotechnica et Informatica*, vol. 12, no. 1, pp. 5–8, 2012.

[10] M. Chau and M. Betke, "Real time eye tracking and blink detection with USB cameras," Boston University Computer Science Department, Tech. Rep., 2005.

[11] F. L. Castro, "Class I infrared eye blinking detector," *Sensors and actuators A: Physical*, vol. 148, no. 2, pp. 388–394, 2008.

[12] A. Dementyev and C. Holz, "Dualblink: A wearable device to continuously detect, track, and actuate blinking for alleviating dry eyes and computer vision syndrome," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 1, pp. 1:1–1:19, 2017.

[13] S. Ishimaru, K. Kunze, K. Kise, J. Weppner, A. Dengel, P. Lukowicz, and A. Bulling, "In the blink of an eye: combining head motion and eye blink frequency for activity recognition with Google Glass," in *Proceedings of the 5th augmented human international conference*. ACM, 2014, p. 15.

[14] M. Kassner, W. Patera, and A. Bulling, "Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*. ACM, 2014, pp. 1151–1160.

[15] M. Pal, A. Banerjee, S. Datta, A. Konar, D. Tibarewala, and R. Janarthanan, "Electrooculography based blink detection to prevent computer vision syndrome," in *Electronics, Computing and Communication Technologies (IEEE CONECCT), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1–6.

[16] M. Zajicek, "Successful and available: interface design exemplars for older users," *Interacting with computers*, vol. 16, no. 3, pp. 411–430, 2004.

[17] S. Ondáš and J. Juhár, "Dialog manager based on the VoiceXML interpreter," in *Proc. 6th Intern. Conference DSP-MCOM, Košice*, 2005, pp. 80–83.

[18] M. Sulír and J. Juhár, "Hidden Markov Model based speech synthesis system in Slovak language with speaker interpolation," *Acta Electrotechnica et Informatica*, vol. 15, no. 4, pp. 8–12, 2015.

[19] J. Juhár, S. Ondas, A. Cizmar, M. Rusko, G. Rozinaj, and R. Jarina, "Development of Slovak GALAXY/VoiceXML based spoken language dialogue system to retrieve information from the Internet," in *Ninth International Conference on Spoken Language Processing*, 2006.

[20] S. Ondas, J. Juhár, M. Pleva, P. Fercak, and R. Husovsky, "Multimodal dialogue system with NAO and VoiceXML dialogue manager," in *8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2017)*, 2017, pp. 439–443.

[21] L. Mackova, A. Cizmar, and J. Juhar, "A study of acoustic features for emotional speaker recognition in i-vector representation," *Acta Electrotechnica et Informatica*, vol. 15, no. 2, pp. 15–20, 2015.