Habilitation Thesis

Multimodal Human Computer Interaction in Specific Environments

Zdenek Mikovec

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Computer Graphics and Interaction
Karlovo nam. 13, 121 35 Praha 2
Czech Republic

Preface

Jestliže něco není dostatečně jasné, nevyžaduje si to učeného sporu, nýbrž důkladnějšího prozkoumání.

If something is not sufficiently clear, detailed exploration is needed rather than learned dispute.

Jan Amos Komenský

In this thesis seven selected research papers I am a co–author of are presented. These papers represent contribution to the research field of multimodal human computer interaction in specific environment. In particular they are focused on the multimodal interaction in mobile environment and on the principles of user interaction design for visually or motor impaired users. Four of the presented papers were published in international journals (three of them are impacted), three other papers were published at established conferences in the field of research.

The thesis starts with an introductory part providing a background information. The introduction is followed by an overview of contributions of the presented papers. The thesis is concluded by a brief summary and ideas for future work in the area. The thesis contains paper reprints in a form they have been published or submitted to the publisher.

Prague, October 30th, 2014 Zdenek Mikovec

Acknowledgments

I would like to thank all my colleagues at the Department of Computer Graphics and Interaction for their cooperation and my family for their support and patience.

Contents

Chapter 1

Introduction

Interaction of humans with computer becomes in last years more and more intensive especially with introduction of mobile, ambient and ubiquitous computing. Computers and their complex user interfaces appear almost in all sophisticated products. This brings both new tasks, which are solved by computers, and variety of new environments where the computers play an important role.

Besides this we can observe massive penetration of computers across the population including also children, elderly and disabled people which increases the diversity of the user preferences and abilities. This brings new challenges for user interface design as these new conditions increase pressure on higher efficiency and intuitiveness of computers. Common user interfaces, where the computer output requires complex visual perception and the user input relies on fine motor activity of hands and fingers, is becoming insufficient. Various and dynamically changing environmental conditions (e.g., weather conditions, audio pollution, fast task switching, occupation of hands due to parallel task performance, occupation of eyes by primary task coming outside of the computer) make this traditional kind of user interaction in many cases almost unusable. Furthermore the emergence of new user groups brings new categories of user preferences and abilities which can be hardly satisfied by common user interaction methods as for example visually impaired users cannot perceive complex visual output, or motor impaired users are not able to perform fine motor movements of hands and finger.

The promising solution to this problem seems to be employment of multimodality in user interface design. Multimodality of user interfaces became a key topic for research of new user interaction techniques especially in context of mobile environments (e.g., outdoor interaction, interaction in cars) or disabled people (in particular visually and motor impaired).

This thesis addresses problems of multimodal human computer interaction from three points of view:

- interaction in dynamically changing environments (Appendix A, Appendix B, Appendix C)
- user interface control (Appendix D, Appendix E)
- text entry (Appendix F, Appendix G)

In the following sections the introduction to above mentioned three problem areas is provided.

1.1 Interaction in Dynamically Changing Environment

Dynamic changes of the environment where the human computer interaction takes place can influence the usability of user interface significantly. Moreover in situations when the user needs to be mobile the environment becomes an integral part of the human computer interaction. Mobile environment brings a broad set of problems for design of user interaction. There can be defined a list of key limitations that need to be taken into consideration by the user interface designer: small screen, slow and imprecise user input, unstable light and noise conditions, high frequency and intensity of external distractors (like occurrence of dangerous traffic situation, interaction with bypassing pedestrians). These characteristics can be perceived as rather static (e.g., screen size) or dynamically changing (e.g, light conditions, frequency of distractors). Existing solutions typically employ unimodal interaction and understand the environment as a static set of parameters. A detailed analysis of mobile environment is often not an integral part of the user interface design process, what results in lower ability of the user interfaces to react adequately on the environment changes.

For complex user interaction like interaction with 3D graphics enriched by semantic information, the unimodal way of information retrieval (virtual walkthrough with use of mouse or keyboard) can not provide us with acceptable results (appropriate level of user satisfaction and usability) in mobile environment. In mobile environment the user often needs to use hands and eyes for other tasks (e.g., interaction with physical objects under investigation). Appendix A addresses this problem by introducing multimodal way of 3D graphics based information retrieval. The semantic description of 3D graphics and detailed analysis of the environment is extensively used in the user interface design process.

In situations where the environment characteristics are changing very dynamically it is necessary to analyze various conditions and understand the user behavior in such conditions to be able to design good user interfaces. Conducting an experiment in realistic environment (field test rather than laboratory one) is complicated both in the observation and analytical phase, what leads to low number of field tests with detailed monitoring of user behavior. In this thesis (Appendix B) an evaluation tool for analysis of the user behavior in realistic environments is introduced.

Very important research field is navigation and orientation of people with limited abilities of orientation in mobile environment. Here I have in mind especially visually and motor impaired people. There is a lot of computer based assistive tools helping disabled people with navigation and orientation. However these tools often either interfere with essential orientation aids (e.g., white cane, sense of hearing) nor exploit the users potential fully, trying to compensate the disability rather than utilizing the user abilities. In Appendix C a study demonstrates unique abilities of visually impaired people when navigating in urban areas.

1.2 Multimodal Control of User Interfaces

Multimodal control of user interfaces can be solved in two ways. First, modalities are used complementary, what means that the user interface must be controlled by more modalities simultaneously. Second, modalities are used supplementary, where the user can choose between more alternatives to control the user interface. The complementary approach typically brings richer interaction which can lead to higher efficiency and usability of the user interface control. The supplementary approach primarily increases the accessibility of the user interface control and reduces the error rate, when using such a user interface.

For users with upper–limb impairment it is typically very complicated or even impossible to control the user interface via direct pointing devices (e.g., mouse, touchpad, joystick). This thesis addresses this problem by presenting a multimodal solution based on head motions and humming, allowing fast and unrestricted control of a virtual mouse (Appendix D).

Very intensive research is focused on user interfaces based on natural language understanding. One of the key problems is the question of effective turn–taking and turn–yielding in dialogs. Humans are very sensitive to this phenomena in natural dialogues and thus it is very important to solve this issue properly. However there is missing knowledge how sensitive the human will be on various cues indi-

cating turn–taking and turn–yielding when an avatar is used in user interfaces. In Appendix E the human sensitiveness on visual and vocal cues for turn–yielding in natural dialogs is analyzed. Based on an extensive user study important insights that help to understand better this problem area are presented.

1.3 Alternative Ways of Text Entry

Entering text is a very complex activity of the user when interacting with a computer. Its efficiency is in many cases crucial for overall usability of a computer. For motor impaired users with limited control of their hands it can be rather challenging to efficiently input a text. There is a need to come up with innovative ways of text entry that must be based on multimodal interaction. One of the field of research in this area is focused on non–verbal vocal input as it is more suitable for motor impaired users (their impairment often affect also the vocal cords) than verbal communication and it is also language independent.

Design of non–verbal vocal gestures is a non–trivial process and with the increase of number of gestures the probability to make a semantic error in the formal gesture description is increasing significantly. There are two most frequent semantic errors. First occurs when there exist two formal gesture descriptions which can be satisfied by one gesture. Second error occurs when a formal gesture description cannot be satisfied by any gesture. Appendix F analyzes the nature of non–verbal vocal pitch–based gestures. Based on a formal description of these gestures a gesture modeling tool was created, which helps designers to avoid semantic errors while defining larger set of vocal gestures.

The efficiency of text entry system based on non–verbal vocal input is dependent on careful design of vocal gestures and on the user interface control mechanism for choosing a letter or string of letters. There are several different layouts of entry systems controlled by limited number of gestures of various nature that show different efficiency. In Appendix G a novel predictive keyboard operated by humming is presented. It was designed especially for motor impaired users who cannot use hands at all for entering text. Four different text entry user interfaces were presented and evaluated in user study which showed interesting relations between the user interface design and the user performance and acceptance.

Chapter 2

Overview of Contributions

In this chapter an overview of main contributions of presented papers is given. Contributions are divided into three sections addressing specific problem areas of multimodal human computer interaction: multimodal interaction in mobile environment, voice–based user interface control, and non–verbal vocal text entry for motor impaired users.

2.1 Multimodal Interaction in Mobile Environment

In the following papers three methods are presented which illustrate a design of multimodal user interfaces based on their exploitation of mobile environment. The first and the third paper focuse on the natural language communication in two different contexts. The second paper shows how important is the ability to analyze the mutual interaction between the user and the environment for design of user interfaces.

Beyond traditional interaction in a mobile environment: New approach to 3D scene rendering. Appendix A investigates the problems of graphical interaction in mobile environment. A prototype of a voice user interface was created that allows users to interact with the graphical data in mobile environment. A method was developed that is able to seamlessly restrict the conversation language used for communication between the system and the user, and thus increase the recognition success rate above 90% what is satisfactory for real life usability of natural language based user interfaces. The restriction of the conversation language is based on analysis of the semantic description of the 3D scene and utilization of the mobile environment the user is currently in. The semantic description is derived

from a domain ontology (e.g., construction site domain) and thus allows generalization and abstraction of the description. In situations where the user context can be properly described (like well–defined work processes) the user will not notice that the conversation language is being restricted.

An evaluation tool for research of user behavior in a realistic mobile environment. Appendix B introduces a tool for analysis of influence of a realistic environment on the user behavior. As the realistic environment (typically field test environment) can influence the user behavior in an unpredictable and mostly uncontrolled way, for researchers, it is challenging to measure and analyze the user behavior. This complex tool provides a unique way of context visualization and synchronization of measured data of various kinds. The synchronization is event based (e.g., start of some task, reaching specific location in the environment), and thus helps to search for interesting behavior patterns important for further investigation. Thanks to this tool it is possible to efficiently evaluate huge amount and heterogeneous data acquired during complex usability tests in a mobile environment. The functionality of this tool is demonstrated on the use case "Navigation of visually impaired users in the building with support of a specialized navigation system". The experiment was focused on collecting and analyzing data that may show level of user stress which may influence user's ability to navigate himself/herself. The analysis focused both on objective data like Galvanic Skin Response parameter, Heart Rate Variability parameters, audio video recordings, observer's logs from the test sessions and subjective data like the users feeling of the stress level. This complex analysis helped us to discover interesting relation between the situation when the user get lost on the route and the ability to remember the route for later navigation.

Collaborative Navigation of Visually Impaired. Appendix C shows that navigation system for visually impaired users can be much more efficient if it is based on collaboration between visually impaired persons and on utilizing distributed knowledge about the environment in which the navigation task takes place. A qualitative study was conducted to gain insight into the issue of communication between visually impaired persons while they are mutually navigating in an unknown environment. Several hypotheses were formulated which were validated by a quantitative study with a sample of 54 visually impaired respondents. A qualitative study was conducted with 20 visually impaired participants aimed at investigating regularly walked routes used by visually impaired persons. The results show that most visually impaired people already collaborate on navigation, and consider an environment description from other visually impaired persons to be adequate for safe and efficient navigation. It seems that the proposed collaborative navigation system is based on the natural behavior of visually impaired people. In addition, it has been shown that a network of regularly walked routes

can significantly expand the urban area in which visually impaired persons are able to navigate safely and efficiently.

2.2 Voice–based User Interface Control

In this section two papers are presented that investigate the potential of voice–based user interface control. In the first paper the efficiency of non–verbal vocal gestures for full–range simulation of mouse is presented. In the second paper the effect of introducing an avatar on turn yielding phenomena in natural language based user interfaces is investigated.

Hands free mouse: comparative study on mouse clicks controlled by humming. In Appendix D a novel hands free method is presented which simulates mouse clicks by humming while the cursor is navigated by head movements tracked by a web camera. This method is based on simple hummed voice commands. It is fast, language independent and provides full control of common mouse buttons. This method was compared with other three different methods in an experiment with more than 50 participants, that showed its efficiency when taking into account task duration. Among hands free methods the humming method was the fastest. However the subjective feel of comfort was the worst, what could be caused by unfamiliarity of the humming interaction technique.

Avatar and Dialog Turn–Yielding Phenomena. Appendix E explores effectiveness of selected visual and vocal turn–yielding cues in user interfaces based on natural language using synthesized speech and an avatar. The aim of this work is to detect the role of visual and vocal cues on dialog turn–change judgment using a conversational agent. The cues were compared and studied in two experiments with more than 70 participants. Findings of those experiments suggest that the selected visual turn–yielding cues are more effective than the vocal cues in increasing a correct judgment of dialog turn–change. Vocal cues used in the experiment showed quite poor results and the conclusion discussed possible explanations of this phenomena.

2.3 Non–verbal Vocal Text Entry

Non–verbal vocal interaction is an interesting alternative way of interaction with a computer. In the following two papers the potential of this interaction method

is investigated. The first paper investigates the problem of design of such vocal gestures. The second paper introduces new text entry methods based on humming.

Understanding Formal Description of Pitch–Based Input. Appendix F presents a tool supporting the design of pitch–based vocal commands (humming, whistling, singing). In previous work a formal description of the pitch–based input was designed, which can be used by user interface designers to define vocal commands. However, as it is discussed in this paper, the formal description can contain semantic errors that are not obvious for the designer. The aim of this paper is to design an efficient method for validation of a formal description with assistance of a designer. This tool is capable of visualizing vocal commands and detecting semantic errors automatically. A user study was conducted that brought preliminary results on comprehension of the formal description by designers and ability to identify and remove semantic errors efficiently. The designers who used the tool were more successful in understanding the formal description and identified more errors.

Humsher: a predictive keyboard operated by humming. Appendix G investigates the potential of non–verbal vocal gestures for hands free text entry. Humsher, a novel text entry method operated by humming was developed. The method utilizes an adaptive language model for text prediction. Four different layouts of user interface were designed and compared. Three of them use dynamic keyboard layout in which n–grams of characters are presented to the user to choose from according to their probability in the given context. The last interface utilizes static layout, in which the characters are displayed alphabetically and a modified binary search algorithm is used for an efficient selection of a character. All interfaces were compared and evaluated in a user study involving 17 able–bodied participants. Case studies with four disabled people were also performed in order to validate the potential of the method for motor impaired users. The average speed of the fastest interface was 14 characters per minute, while the fastest user reached 30 characters per minute. Disabled participants were able to type at 14 – 22 characters per minute after seven sessions. The study showed that humming is a reasonable text entry method for motor impaired people.

Chapter 3

Summary and Future Work

Papers presented in this thesis try to clarify several aspects of multimodal human computer interaction in specific environments. Contributions are focused on following three areas: multimodal interaction in mobile environment, multimodal control of user interfaces, and alternative ways of text entry. Several new interaction methods and analytical tools were developed, and extensive quantitative user studies were conducted gathering evidence for validation of hypotheses of user behavior in specific environments.

It was shown (Appendix A) how the recognition ratio of natural language based user interfaces can be increased by exploitation of a user context. Usage of sophisticated evaluation tool (Appendix B) can help to discover interesting behavior patterns (Appendix C).

In Appendix D it is shown that development of new method of UI control based on vocal gestures can lead to more efficient user interaction. By performing extensive user study (Appendix E) it was analyzed the role and importance of visual cues for control of natural language dialogs.

A tool for design of vocal gestures (Appendix F) increased the efficiency of design of multimodal user interfaces. By designing new text entry method based on vocal gestures (Appendix G), a usable alternative text entry system was designed for motor impaired users.

Future research in the area of multimodal user interaction can be directed towards creation of a dynamic model of user interaction abilities. It was observed that restriction of some interaction ability (given by environment conditions or impairment) is followed by improvement of another existing ability (e.g., hearing of blind person) or development of a new one (e.g., reading text via haptic modality).

This is possible due to neural plasticity mechanisms influencing the functions of brain and the learning effect. There are currently several theories how the neural plasticity influences the brain. Enriching the model of user abilities by findings from these theories can introduce a dynamics into the model which is essential for modeling of impaired or elderly people, where the abilities changes significantly over relatively short period of time. Such dynamic model can support design of new interaction techniques that will much more efficiently utilize the interaction abilities of a user in specific environments. This model can control automated generation of user interfaces, what can be an important step towards user interfaces that are able to adapt to the user continuously and seamlessly as the abilities of the user changes in time. Another interesting research direction can be investigation of how the stress influences the cognitive processes and ability of the user to interact via various modalities with a computer. With detailed knowledge of these effects it can be possible to design multimodal user interfaces that will be usable in wide range of conditions with different stress level or to introduce mechanisms for automated adaptation of user interfaces to dynamically changing stress conditions.

Appendices – Paper Reprints

Appendix A

Beyond traditional interaction in a mobile environment: New approach to 3D scene rendering

Mikovec Z., Cmolik L., Slavik P.: Beyond traditional interaction in a mobile environment: New approach to 3D scene rendering. In Int. Journal of Computers & Graphics. 2006, vol. 30, no. 5, p. 714-726. ISSN 0097-8493. **IF=0.794**

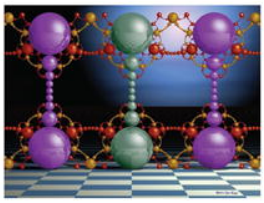
Volume 30 Issue 5 October 2006 ISSN 0097-8493

COMPUTERS & GRAPHICS

An international journal of systems & applications in computer graphics

Algorithms and techniques for interaction, multimedia, modelling and visualization

EDITOR-IN-CHIEF: J. L. Encarnação



In this issue the special topic is
MOBILE COMPUTING AND AMBIENT INTELLIGENCE
Guest Editors: Heidrun Schumann, Thomas Kirste

Available online at www.sciencedirect.com
ScienceDirect
Computers & Graphics 30 (2006) 714–726
COMPUTERS & GRAPHICS
www.elsevier.com/locate/cag

Beyond traditional interaction in a mobile environment: New approach to 3D scene rendering

Zdenek Mikovec*, Ladislav Cmolik, Jiri Kopsa, Pavel Slavik

Czech Technical University in Prague, Karlovo Namesti 13, 121 35 Praha 2, Czech Republic

Abstract

In this paper the problems of user interaction in a mobile environment are investigated. Interaction in this environment imposes new requirements both on UI designers and the users. This paper deals in particular with problems of graphical interaction in a mobile environment. The interaction in a mobile environment requires new approaches that will result in a new type of rendering of graphical data. A rendering technique that enables rendering and annotation of objects in a 3D scene on mobile devices is presented. This technique is based on transformation of the 3D scene to a 2D vector graphic representation. The input 3D scene is given in the VRML format and the output 2D format is SVG. In the second part of this paper we are presenting a prototype of a voice user interface that allows users to interact with the graphical data in a mobile environment. The semantic description of the scene plays the key role in restriction of the language used for communication. The communication between the user and the mobile system is performed by means of natural language that is restricted according to the context the user is currently in. The implementation of the voice user interface is based on the existing VoiceXML platform.

Keywords: Mobile computing; User interfaces; Graphical interaction; Voice recognition; Ontology; XML; SVG; OWL

1. Introduction

In this paper we will deal with problems of user interaction in a mobile environment. Interaction in this environment imposes new requirements both on UI designers and the users. We will focus on interaction with application data that have graphical flavor (like blueprints, plans, etc.). The interaction with this type of data has been performed for decades in a static environment (e.g. PC, characterized by large screen, mouse, keyboard) for which some feasible interaction principles were developed. The interaction in a mobile environment requires new approaches that will result in a new type of rendering of application data.

These new requirements stem from several new issues typical for mobile computing: small screens, context dependency, new interaction devices like a stylus, dynamically changing environments, often task switching based on external events, etc. Such an environment will impose significant limits on interaction as the user's comfort when using mobile devices is much lower than in a static (e.g. PC) environment. If we focus only on the modification of the user interface we will very quickly encounter limits determined by the application data presented. To get beyond these limitations and reach much higher adaptability of the interaction we need to introduce new methods for application data rendering and interaction.

The rendering can be in a broader sense understood as a way to present application data to the user, by means of targeting different human sense organs. In our case we focus on two senses: seeing and hearing and thus we can divide the presentation modalities to visual and audio ones (see Fig. 1). The visual one will be based on traditional graphical rendering methods, which will be modified according to the situation in a mobile environment. The audio modality in our case will be text based, where by means of verbal communication it will be possible to get information about the structure of the scene. In this case

*Corresponding author. Tel.: +420 22435 7647; fax: +420 224 923 325.
E-mail addresses: xmikovec@fel.cvut.cz (Z. Mikovec), cmolikl@fel.cvut.cz (L. Cmolik), j.kopsa@fee.ctup.cz (J. Kopsa), slavik@fel.cvut.cz (P. Slavik).

0097-8493/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cag.2006.07.022

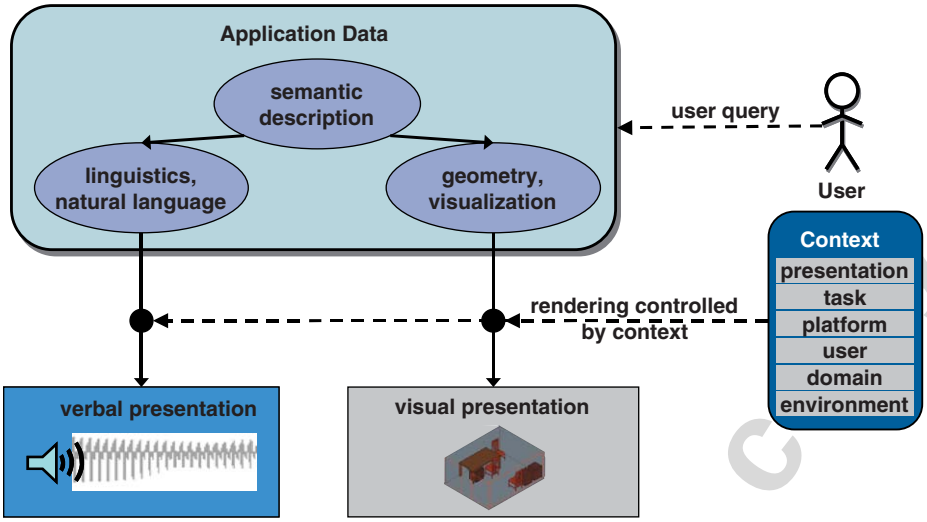


Fig. 1. Two types of rendering of graphical data.

the user formulates proper queries and the system answers in natural language.

The interaction with application data is in our case understood as a solution of navigation problems in 3D graphical data. We also assume simple manipulation with the data (e.g. creation of annotation to an already existing object).

For satisfactory adaptability of the rendering and interaction it is necessary to get information about the structure and the meaning of the data (3D scene in our case). To achieve this we introduce a semantic description of the application data, which is the starting point of both of our rendering methods (see Fig. 1).

Our goal is also to allow natural mixing of both types of rendering driven by the current user context (see Fig. 1). It is obvious that various aspects of context will influence the way application data are rendered and the corresponding interaction. The context in general imposes various constraints on user activity. These constraints stem both from limitations on the side of a device (like a small screen) or on the side of the user (some kind of impairment) or they can be the result of a specific situation the user is currently in (like being in a certain floor of a large building where the inspection is performed—see Section 2). All these constraints may in general influence the method of interaction with application data, thus influencing the particular way of data rendering (e.g. defining the user's viewpoint in the 3D scene) and the user interface by means of which we interact with graphical data. As the problem is rather general we will concentrate on specific issues that will show our approach to the solution of some aspects of this general problem. In our research we concentrated on development of new methods for scene rendering and interaction with such a scene.

In Section 3 of this paper we will discuss modification of the graphical rendering process (adjusted to the needs of a mobile computing environment) [1]. In Section 4 we will discuss a new approach to textual based rendering that could be an alternative to graphical rendering in a specific context the users might be in [2]. Under textual rendering we will understand the process where the users acquire the information about the structure of a 3D scene by means of textual based interaction (queries and answers are in textual form) in audio modality. Both approaches have been tested on a number of examples and the results were very encouraging.

2. Use case

Our use case presented below is derived from the facility management (FM) application domain. In the FM domain the workers typically work on move. While walking through different buildings they perform various tasks like checking the inventory, finding hazardous materials, etc. In our particular use case the worker is called inspector. The inspector's task is to match the inventory in the building with the information stored in the inventory system (see Fig. 2). The inspector has at his/her disposal an inventory system on a mobile computer (like a PDA, or TabletPC). The inspector enters the office room and starts the inspection by asking the system verbally to list the inventory items in the investigated room (in this case an office). The system also answers verbally by listing the inventory numbers of all items stored in the system of the office room. The inspector finds some discrepancies (like wrong inventory number, missing item) and corrects them immediately (still in voice modality). This audio modality is very convenient for the inspector, because the inspector has free hands for manipulating the inventory items and is not limited when reading the inventory numbers. After this activity the inspector needs to do a deeper check of the types of inventory items (like the exact position, shape and material of the items). For this task the inspector switches to the visual modality, where the system offers a 2.5D view of the office room. The inspector can navigate in the visualized scene and finds that one container has a different

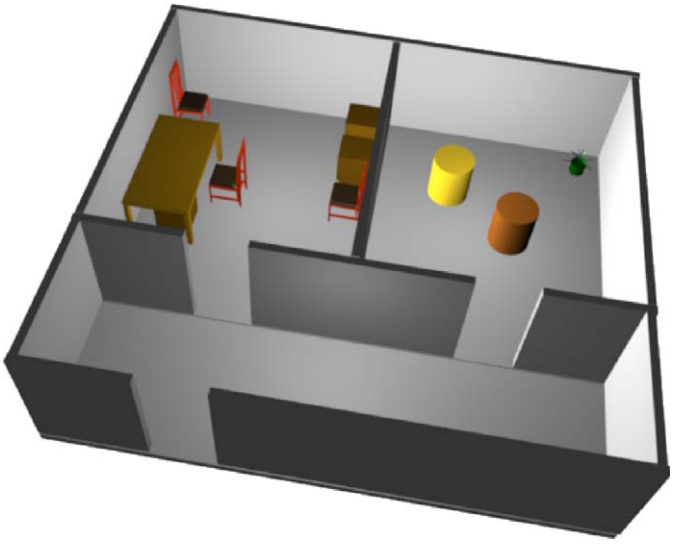


Fig. 2. A 3D plan of the building floor inspected by the inspector.

height (different number of drawers). The inspector makes a quick short annotation to the container in the scene noting this error. This annotation is done in audio modality.

As shown in our use case for efficient task solving the inspector needs to have specific tools at his/her disposal to investigate, annotate and interact with the application data in a mobile environment.

The tools must provide information in two modalities (visual—in a form of building plan; audio—in a form of structured textual description of the building and its inventory) to increase the efficiency of the inspection. The switching between these two modalities must be possible at any time.

On the presented use case we can see that the tools should have a more flexible user interface (multimodal one) and should be less demanding from the point of computational power (as mobile devices and wireless networks are much less powerful than standard PCs and LANs).

3. Graphical based rendering for mobile environment

The interaction with 3D scenes in a mobile environment is very difficult. The problem of navigation in the 3D scene is especially noticeable. In a static environment the navigation requires concurrent usage of special keys to switch between navigation modes (move, pan, rotate, fly, walk, etc.), in combination with a pointing device (mouse, joystick, etc.). Mobile devices do not allow us to perform such navigation comfortably. Moreover, common users are not trained to work in the 3D environment with so many degrees of freedom, which results in the loss of orientation in the 3D scene. As mentioned in [3] users prefer a 2D interaction environment or some kind of combination of the 2D and the 3D environment. The 2D interaction environment reduces the user interaction to two basic functions—zoom and pan, which is a much safer environment, where the user in general does not lose orientation and can navigate much faster.

There exist several approaches to the rendering of 3D scenes on mobile devices, which can be divided into three categories:

- Server-side methods. These methods [4] provide remote rendering of the 3D model as a reaction to user-interaction-requests from the mobile device and send rendered raster images back to the mobile device.
- Client-side methods. These methods [5] render the 3D scene completely on the mobile device.
- Hybrid-side methods. These methods [6,7] balance the workload between the server and the client to decrease network communication and improve performance on the client side.

All of the solutions mentioned above focus only on the rendering or on the load distribution between server and client side. None of the mentioned works focuses on the user navigation and interaction nor takes the navigation or interaction into account.

3.1. Our approach

In our solution we have started with the analysis of user needs. As mentioned before, the users prefer the safer 2D interaction environment or a combination of the 2D and 3D interaction environment. Therefore, we introduce a solution that allows the users to work in the 2D interaction environment while preserving their illusion of being in 3D.

Our approach is a hybrid-side method. The distribution of the workload between server and the client (mobile device) plays a very important role in our solution of the rendering problem in the mobile environment. Our approach is based on transformation of the boundary representation of a 3D scene by means of a projection to the 2D vector image (see Fig. 3). The transformation preserves the object oriented representation of the 3D scene—each 3D object can be wholly (all faces) identified in the 2D vector image. The resulting 2D vector image has 2.5D representation (each 2D object is located in its own projection plane and these objects are ordered by their distance from the camera). In other words, each face from the 3D scene has its representation in the 2D vector image (see Fig. 4). This object oriented approach allows the user to interact with the objects in the 2.5D representation.

The scalable vector graphics (SVG) format [8] was chosen to implement these features. SVG is an XML-based format developed for description of 2D vector graphics. It is object oriented and supports zooming, panning and interaction with the objects. Each SVG conformant viewer implements the painter's rendering model [8]. An object defined later in the SVG file is painted over objects defined before. This could be interpreted as the objects defined later in the SVG file are nearer to the viewer than objects

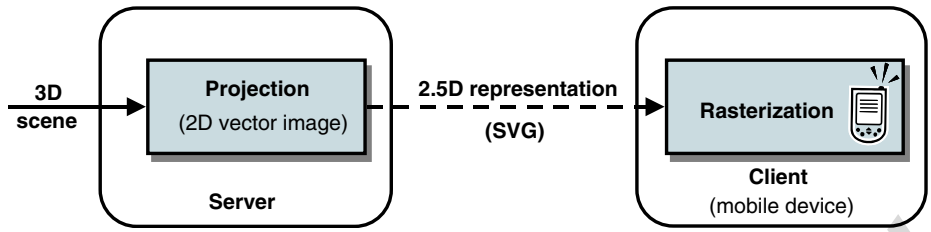


Fig. 3. Distribution of the rendering process of 3D scene.

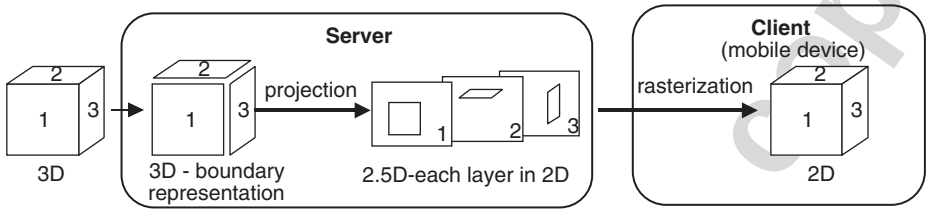


Fig. 4. Projection of a 3D object into the 2D vector image. Each face of the 3D object is projected into a separate layer in 2D vector image. This representation is in fact a 2.5D representation.

defined before. Thus, we can project each face in a 3D scene into a separate layer and use the SVG as the 2.5D representation format. Only the ordering of the faces is known, but the real distance of faces from the viewer is unknown.

It was necessary to develop an algorithm that sorts the faces in a 3D scene in such a way, that a proper image is obtained if the painter's rendering model is applied to them. Such an algorithm is based on the priority list approach introduced by Newell et al. [9]—also known as the painter's algorithm. In our approach the painter's algorithm generates an ordered list of 2D projections of 3D faces. Each individual 2D projection is in vector representation (SVG) and thus the whole SVG file is for the current view a 2.5D representation of the 3D scene.

The transformation of a 3D scene to 2.5D representation is performed on the server side by a modified rendering pipeline. The structure of our modified rendering pipeline is a standard one, except the back-faces are not culled and the solution of visibility is performed in the object space by the painter's algorithm. All parts of the algorithm are modified to be able to handle back-faces. The back-faces are preserved to allow various rendering modes. The 2.5D representation of the 3D scene is sent to the mobile device, where it is rendered. The rendering of 2D data is computationally less expensive than the rendering of 3D data and therefore the response of the mobile device on user interactions is faster.

3.2. Various rendering modes of projected data

The painter's algorithm preserves the information about all objects and also about their ordering with respect to the camera view. Investigation of the 3D scene in 2D space does not allow the user to access all the 3D information (like information about an object hidden behind another one). Our approach reduces this problem by benefiting from the 2.5D representation and by introducing special modes of object rendering.

The SVG used for 2.5D representation of the 3D scene is an XML-based format, therefore the CSS (Cascading Style Sheets) can be used to define various rendering modes of the 2D objects. Thanks to our modified rendering pipeline, which preserves back-faces, we are also able to introduce wire-frame and semitransparent rendering modes. To each SVG file a CSS representing the rendering modes is attached. Moreover, SVG enables interactive and selective change of a rendering mode of an object in the SVG file. This support of various rendering modes compensates in some way the movement of the avatar in the 3D scene. The virtual walkthrough is replaced by features allowing the user to investigate the structure of the 3D scene. We will demonstrate this feature on the earlier described use case (see Section 2).

3.3. Use case

Let us consider an inspector equipped with a mobile device (e.g. PDA) visiting an office (having a 3D model on the server). The task of the inspector is to revise the facility of the office. The inspector retrieves the information needed from the server through the wireless network. The inspector uses a 2.5D environment to investigate the office. The office was selected based on the current context of the inspector. In our use case the inspector requests the 3D data by uttering the following request: "Show me a 3D visualization of the office!" (see Section 4). The data delivered to the PDA are not in the original 3D form (that is not suitable for use in a mobile environment) but they have been transformed to 2.5D representation by means of the process described above. For each room four 2.5D images are created. Each image defines one view of the

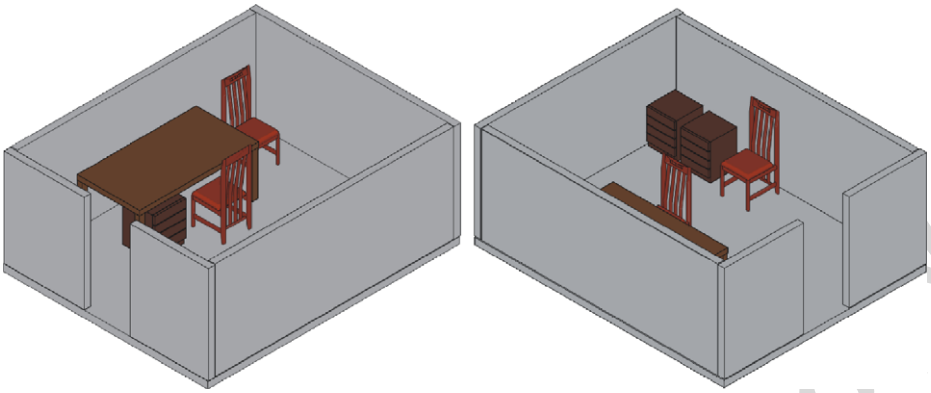


Fig. 5. Examples of initially rendered 3D scene (the office room from two different viewpoints).

room (see Fig. 5 for two of them) form different viewpoint. This allows the inspector to partially rotate the room by selecting the proper view.

The inspector can view the 2.5D representation in various rendering modes (the objects can be displayed in solid, semitransparent or wire-frame mode). Moreover, zooming and panning can be used without sending any request to the server. The inspector can also annotate the objects in 2.5D representation to record discrepancies found (in comparison with reality) in the data.

The inspector needs to examine the scene in detail. The 2.5D representation in Fig. 5 is not informative enough, because some objects can be hidden behind other objects. The inspector changes the rendering mode of the front walls to wire-frame mode to see what objects are hidden behind them (see Figs. 6 and 7).

By using the zoom function the inspector can get closer to the hidden objects and inspect them. While the objects are projected with parallel projection and described in 2D vector graphic the rendering quality during the zooming is preserved and from the inspector point of view it looks exactly the same as zoom in the original 3D scene.

Now the inspector has found out that there is a container with four drawers below the table, but in reality there is a container with three drawers. Therefore, the inspector wants to attach an annotation to the container. For annotation purposes s/he zooms the container in order to create an unambiguous relation between the annotation and the container (see Fig. 8).

Now the inspector can create a multimedia annotation, e.g. voice record and attach it to the container (see Section 4).

Notice that the whole process of scene exploration is done only on the client side (no requests to server are sent). The request is sent only in the phase of annotation creation or when the viewpoint is changed.

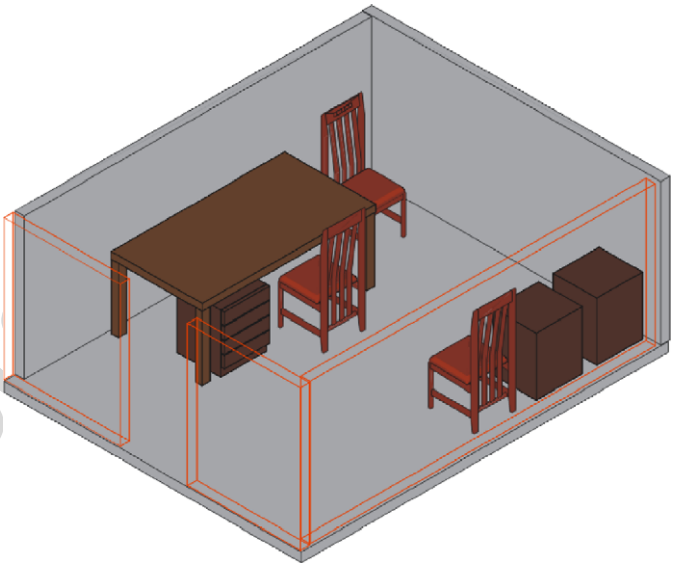


Fig. 6. The front walls rendered in wire-frame mode. The furniture hidden behind the walls could be identified.

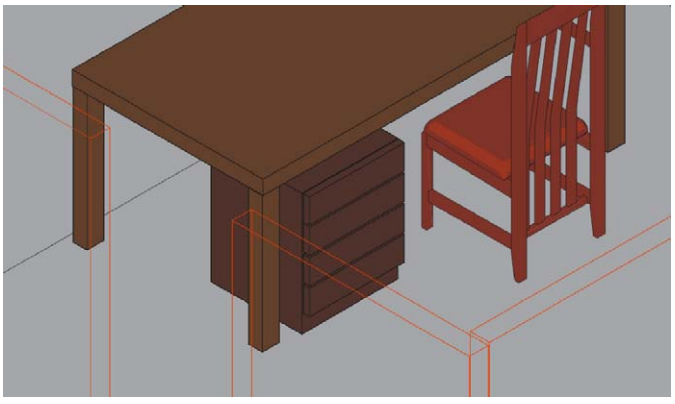


Fig. 7. Inspector performs zooming function to investigate the objects.

3.4. Testing of rendering method performance

This testing was focused on the performance of our application. The performance test was focused on solution of visibility where our modified painter's algorithm spends a significant amount of time from the whole time spent on the rendering. Our implementation [19] was tested on various 3D models of different complexity (number of

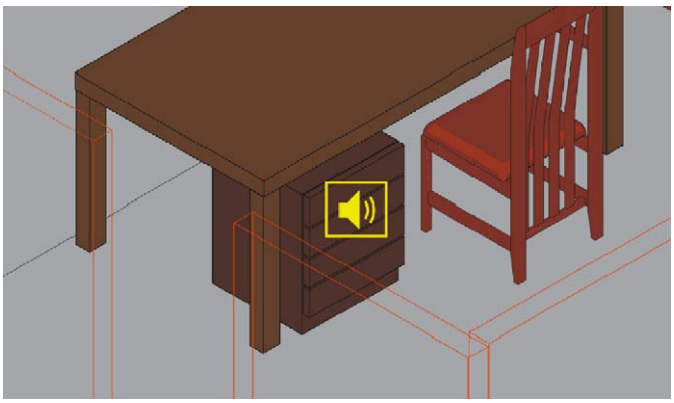


Fig. 8. The inspector has created an unambiguous relation between the container and the annotation.

model	# of surfaces	visibility time [ms]	file size [B]	compresed file size [B]	compression ratio
basic shapes	628	230	118 657	15 320	7.75
wheel	928	611	194 448	83 658	2.32
blender monkey	946	490	183 925	25 148	7.31
holes	1 058	581	212 150	26 310	8.06
torus knot	1 440	771	262 420	33 764	7.77
blender monkey 2	2 032	1 301	424 641	62 593	6.78
stanford bunny	2 915	2 593	536 186	67 053	8.00
pillar	5 260	6 590	953 910	130 415	7.31
cow	5 804	5 668	1 098 199	141 012	7.79
blender monkey 3	8 128	13 529	1 644 489	239 314	6.87

Fig. 9. Performance time of visibility computation and file size reduction after gzip compression.

surfaces). The time spent on the painter's algorithm and the size of the SVG file was measured. The data were measured on a PC equipped with Pentium 4 Mobile running at 1600 MHz and with 512 MB of RAM. The results are presented in the table in Fig. 9.

The size of the SVG file depends linearly on the number of faces. The computational complexity depends quadratically on the number of faces (see graph in Fig. 10). This corresponds with the $O(n)^2$ computational complexity of the painter's algorithm. The measurement showed that the performance of our implementation of the painter's algorithm is sufficient for scenes with a relatively small number of faces (up to 2000). In such a case the time of 2.5D representation generation requires about 1 s. The typical scene used in our use case contains up to 1000 faces. These scenes can be easily handled in a mobile environment.

We should also mention that the size of the SVG file is always larger than the size of the file describing the corresponding 3D scene in VRML format [10]. The SVG file contains much more detailed description of the scene, e.g. the color is stored in the SVG file for every face. In VRML usually one color is defined for the whole object. The size of the SVG file does not create any problems during data transmission as it was compressed in our experiments with compression ratio from 6.78:1 to 8.06:1.

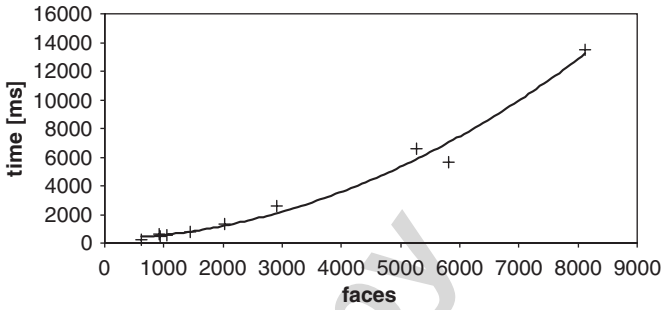


Fig. 10. Performance of our painter's algorithm in dependency on number of faces.

3.5. Usability testing

This test was focused on the approval of our assumption that our approach leads to simplification of the user interaction (by introducing a 2D environment for manipulation with the 3D scene) while the feelings of the user will be as close as possible to a true 3D scene exploration (by introducing the 2.5D representation and rendering modes).

We performed usability tests with six selected users on two different scenarios. All users were exploring the same scene, but half of them in a 3D environment and half in our 2D environment.

The first scenario was similar to the use case described in Section 3.3. The task was to explore the given scene and to check for all objects in the scene. The test results showed that the users had much less problems with navigation in our 2D environment. Moreover after a short time of using our system they forget that they are only in 2D space, which proved that our system induced an illusion of being in 3D space.

In the second scenario the users were exploring one street of a city (see Fig. 11). Their task was to perform a short walkthrough in the scene and again to check for all objects in the scene. The results are similar to results of the first scenario. The users exploring the scene in our 2D environment had again much less navigation problems than users exploring the scene in the 3D environment. A very interesting result of the second scenario is that users exploring the scene in our 2D environment used mostly the zoom function to move backward and forward and suppressed usage of the panning function.

4. Textual based rendering

In a mobile environment the users often get into a situation where they need eyes and hands free for performing their task (e.g. carrying a bag, taking samples of materials with gloves on their hands). At the same time they would like to interact with their mobile device to browse through some information or to make an electronic note. It is obvious that the common way of interaction with the mobile device (using stylus with touch screen and watching the display) becomes unusable in such a situation.

Fig. 11. Walkthrough in the 3D scene. The user is exploring the 3D scene in a 2D environment. Due to the perspective projection and vector nature of the graphics the zooming action is perceived by the user as walking on the street.

We have to switch to different interaction modality—to an audio one. Unfortunately, the existing solutions for audio interaction face serious usability issues when trying to introduce natural language interfaces. Especially in a mobile environment, where the end-devices have low computational power and lots of disturbing noises occur, the ability of such systems to understand the user queries decreases dramatically and the system becomes unusable.

The success rate of understanding user queries is often improved by restricting the language to a specific application domain [11,12]. In a broader sense contextual information is used in most of the voice applications to cope with the restriction process. The contextual information is obtained from various sources (sensors of user gestures, environment sensors) and used to restrict the language as well as to resolve the semantic ambiguity of the natural language as described in [13–15]. This approach does not provide the system with sufficiently detailed context information. There exist several approaches like [16] which try to build up more general multimodal interfaces for multiple applications. These approaches are based on the semantic description of the application domain (by means of ontology), which should be automatically connected with the generic multimodal interface.

In our case we need to interact with graphical data by means of audio modality. This leads to introduction of textual based rendering methods, which will allow verbal communication. To offer the user a comfortable natural voice interface, the communication language starts to be very complex too. Such a complex language cannot be satisfactorily handled by the general voice recognition system.

4.1. Our approach

The large size of the language needed to describe complex graphical data is the main issue. Our solution is based on typical use case in a mobile environment, where the user is moving from one place to another and solving problems related to those places. In our particular use case (see Section 2) the user is an inspector equipped with a mobile device inspecting the facility of the warehouse. The conversation context can be defined with respect to the inspector's context (environment, current task, etc.—see Fig. 1). This conversation context is used to reduce the size of the language to a subset, which will be sufficient for the expected conversation.

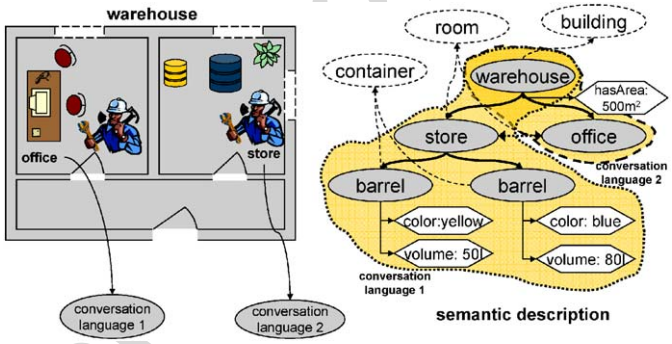


Fig. 12. Use case of construction site inspection: the inspector location restricts the conversation language. The language is derived from an appropriate part of the semantic description.

Fig. 12 explains the approach presented. When the workers are in the store of the warehouse, they will most probably ask about the barrels that are placed in that room or about objects that are related to them. Based on the workers location we generate restricted conversation language, which covers only the most probable discussion topics (see "conversation language 1 and 2" in Fig. 12). The probability that the workers will ask about the objects in the office is significantly lower in the given context. If the inspector asks about objects that are out of the current restricted conversation language, the system will not understand and remind the inspector about the current conversation context (e.g. location of the inspector).

The novelty of our approach is that the conversation language is defined by both contextual information (described in Fig. 1) and the semantic description of application data (see Fig. 12)—in our case the construction site plan—which is the subject of conversation. By introducing the semantic description we obtained much more detailed information about the conversation context. We are able to dynamically and seamlessly (from the user's point of view) restrict the conversation language according to the real conversation context. In every particular moment of the conversation only a relevant subset of the language is used (in other words: a specific context oriented language will be used). The union of these particular languages represents a general language large enough to

cover the general conversation about the given topic (in our case the union of all restricted languages covers all the rooms in the building and activities that could be performed there).

Besides navigation in the scene our approach also solves the problem of manipulation with the 3D scene. This manipulation includes both annotation and modification of the scene. In our system we understand that annotation is the possibility to create a multimedia content (e.g. voice record in WAV format, photo in JPEG). Modification might concern both geometric parameters and a semantic description (related to application domain) of the scene. For annotation there exists a set of commands the user has at his/her disposal.

The whole conversation with the system is led by user-initiative conversation; the user forms queries and commands and the system answers by providing requested information about the scene or performing the command.

In the following text we will describe our approach as a text-based communication between the user and the system. Formulation of queries and the system response will be handled on the textual level. In a mobile environment textual communication is not an adequate one, so the next step was the conversion of text-based communication into a voice-based one. This conversion was realized by means of the voice recognizer developed by IBM and an open-source voice synthesizer.

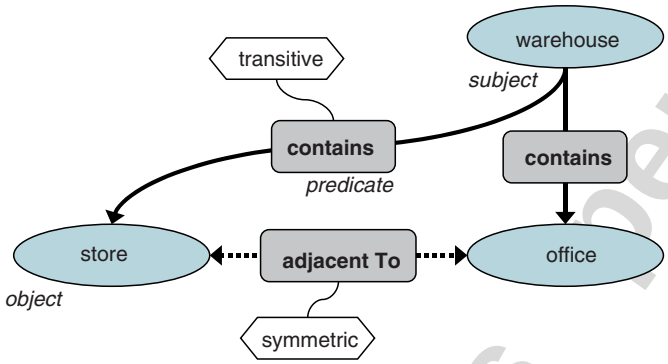


Fig. 13. Ontology structure—triplets: subject–predicate–object.

4.2. Semantic description—ontologies

Formally, the ontology used for semantic description consists of elements, their properties and relationships between them. The ontology is represented as an oriented graph with elements as nodes and relations as edges. An edge and its two nodes represent a fact in the form of a triplet "subject–predicate–object" (see Fig. 13). For example, the relation between the nodes "warehouse" and "store" represented by the predicate "contains" describes the following fact:

warehouse contains store

Element or relation properties may be of different types (property type) like transitive or symmetric. The transitivity of the relation (predicate) "contains" may result in the following fact (see Fig. 14):

warehouse contains barrel with inventory number 1

From this example we can see how the ontology is structured and interpreted.

The application data (3D scenes in our case) are semantically described in textual form—as an ontology (OWL [17]), which can be perceived as an oriented graph. The semantic description is created manually during the authoring process. In this case the textual based interaction is the solution to the problem given above. An ontology example (where combination of graphical data and its semantic description is stored) is presented in Fig. 14.

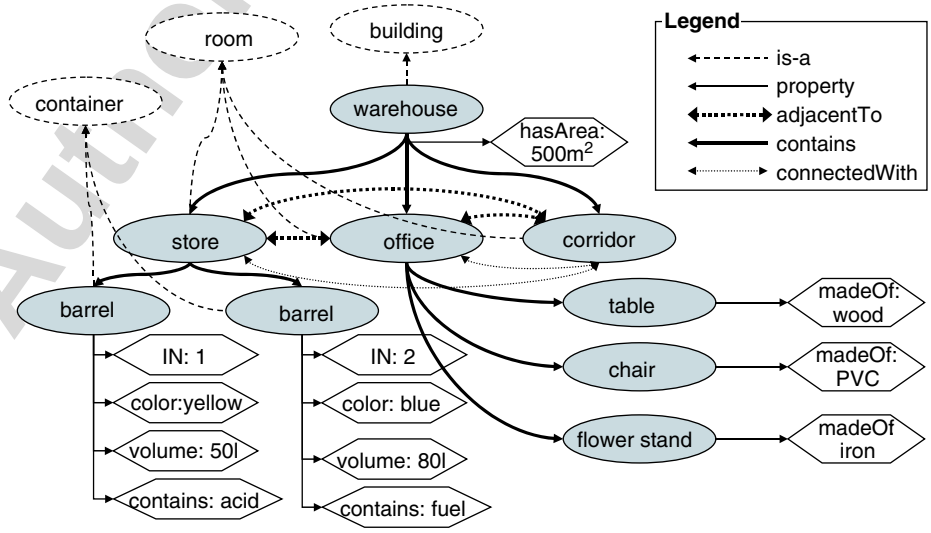


Fig. 14. Semantic description of application data (construction site shown in Fig. 12).

The semantic description consists of a domain and application specific description. The domain description defines abstract terms, classes of objects and their relations (marked by dashed lines in Fig. 14), while the application specific description specifies objects describing a particular construction site or facility. In Fig. 14 the nodes and edges marked with dashed line represent the domain description and the remaining nodes and edges represent the application description. For example, the room node has more general meaning than the store and office objects (the dashed edges represent links between abstract and concrete entities).

4.3. Conversation

The conversation from the user's point of view consists of formulating queries or commands. The commands are added to the conversation language based on the task model of the user (e.g. switching between graphical and audio modality; creating voice annotation for an existing object). As outlined above, the user can also formulate queries to ask about the facts (semantic description) stored in the semantic description (Fig. 14).

The key feature of our solution is that the contextual information is integrated with the semantic description. All necessary contextual information in sufficient detail is available to our system. Because of the linguistic markings made to the semantic description it is possible to access the application data in a natural language (in our case voice-based communication).

There are usually several forms of the same query (the use of synonyms). The query is specified with two key items—source set and target—and several other less important attributes.

The source set represents a node or set of nodes whose properties the users are interested in. The target specifies the properties, which the user is interested in. The two synonymic queries in the following example have the same meaning. Its source set contains one object—*store*—and the target is the property *contains*.

H *(Human)*: What is contained in the store?
OR
H: List the contents of the store!

The corresponding query is executed and the result of its execution is formed into an answer. The system would respond in the following way:

C *(Computer)*: It contains two containers.

For the construction of this answer the generalization stored in the semantic description is used. The abstract terms (like container) are defined as a general term for two barrels located in the store (see Fig. 14). This abstract term is used for constructing the answer given above. The plural form of the word container is defined by the linguistic marking. In the process of the query execution the conversation context is updated. Then the conversation language is modified based on the updated conversation context. The conversation context holds the state of the conversation. This conversation state is defined by the current user context, pointer to the objects in the ontology and conversation history—see Fig. 15. Except in the restricted language production, the conversation context is also used to resolve query ambiguities. For example, the identifier *store* used in previous user queries may be ambiguous since there may be a number of objects with the name "store". However, it can be resolved according to the current user location contained in the user context.

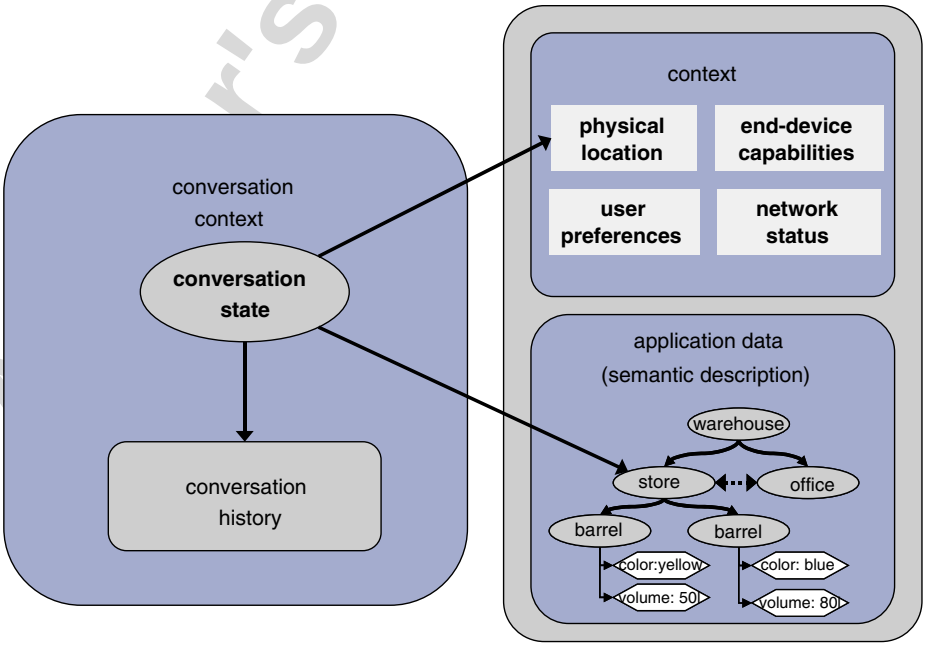


Fig. 15. Conversation context.

The conversation history is a queue that contains references to recently discussed objects of the ontology. It is used to resolve ambiguities of the natural language. For example, the conversation may continue in this way:

H: What is its area?
C: Its area is 500 m^2.

The pronoun *its* is resolved to object *store* by searching the conversation history. The notion of the conversation context is shown in Fig. 15.

The user can also create objects by forming special commands:

H: Add new sample 123 to barrel with inventory number one.
C: A sample 123 related to barrel with inventory number one was added!

This is aimed to fulfill the use case of taking samples—the inspector wears gloves and takes samples with tools. At the same time the inspector creates voice annotation describing the process of taking samples. The physical sample and the annotation are interlinked via the unique identification (in our case "`123`").

The range of allowed queries and commands is quite broad, e.g. the user may also specify a condition that must be met for all objects that are included in the answer:

H: List containers contained in the store and manufactured by Liquids Ltd.

For the switching between the two modalities (textual one and the graphical one) there is a set of commands. One form can be as follows:

H: Show me a 3D visualization of this room.

The detail specification of the whole language along with its semantics, query execution, and answers formatting processes can be found in [18].

4.4. Toward the natural language

The main task when designing the natural language UI was to bridge the gap between the application data and the natural language. The application data described by the ontology does not contain information essential for the system to understand the user queries formulated in a natural language. Also without this information it is impossible to generate system answers to the user in a natural language as well. We have introduced linguistic patterns for generation of sentences from application data in a natural language (see Fig. 16). These linguistic patterns are special attributes added to the abstract level of the application data description (domain description—see Section 4.2). These attributes are then used for generation of the linguistic markings.

The linguistic markings play the fundamental role in the sentence generation process. These markings are generated by the creator of application data or automatically from the domain description linked with ontology elements. These linguistic markings control the process of creation of sentences in a natural language. During the control process the words are modified and placed in a proper place in the sentence in accordance with grammatical rules of a given natural language. The structure of ontology that allows us to analyze and generate sentences for a class of scenes is quite general.

4.5. System architecture

Text-based interaction performed by means of manual input is not suitable for a mobile environment. That is why the user input in a mobile environment should have the form of voice input. We have used the existing VoiceXML server platform for speech recognition and synthesis. The voice recognition part has been developed by IBM. It is configured to request VoiceXML documents from our system, which implements the interaction logic, conversation state management and application ontology data retrieval. The mobile devices are connected to our system with voice-over-IP clients. The architecture of the system implemented is shown in Fig. 17. Our system that we implemented fulfills both requirements for flexible communication between server and client and the easy conversation between the user and the system in a natural language.

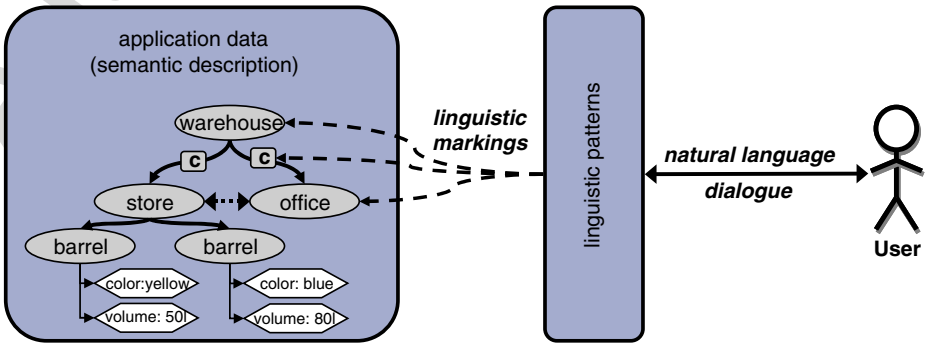


Fig. 16. Linguistic patterns used for transformation of application data into natural language.

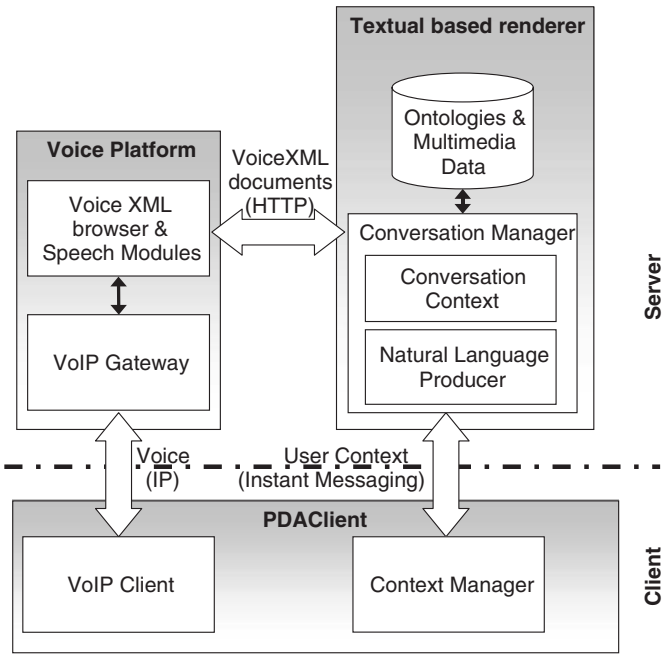


Fig. 17. System architecture.

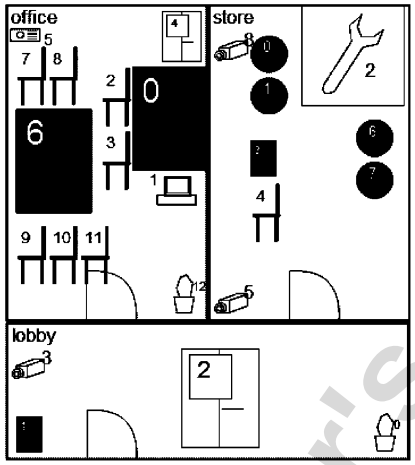


Fig. 18. The environment setup of the usability test.

4.6. Testing our approach

In this paper we have presented an ongoing work, but we are still in the process of development. We have implemented a prototype for testing purposes. The user acceptance of this communication method was tested by means of several usability tests. Our usability tests were focused on two aspects: first, to determine the size of the input language where the speech recognition system is reliable on the required level, second, to test our hypothesis that during the conversation we are able to dynamically reduce the input language (with respect to the current user context and detailed application ontology description) in such a way that the users will not be restricted in their queries (see Sections 4.1and 4.4) (Fig. 18).

The usability test was performed with 12 users. The user's task was to do an inspection in several rooms of a building. They had to check the objects in those rooms and make voice annotations where necessary. During the test we have determined the size of the input language where the speech recognition system was at least 90% successful in recognizing the user queries. The conversation context given by the user context, application ontology and conversation history was continuously changing and based on its current state the input language was dynamically generated to allow the natural language conversation with the user.

5. Conclusion

In this work new interaction methods for manipulation with 3D scenes suitable for a mobile environment have been presented. These methods are based on changing the rendering process of the application data to allow much higher adaptability of the whole interactive system to the user needs.

The first method is focused on modification of the graphical rendering process. The rendering process is distributed between the server and the mobile device. The 3D scene is projected to 2.5D representation on the server side and this 2.5D representation is delivered through the network to the mobile device. The users can interactively change rendering modes of objects in the 2.5D scene (e.g. solid, semitransparent or wire-frame mode). The users can also zoom and pan the 2.5D scene. Moreover, they can annotate the objects in the 2.5D scene. The possibility to choose various rendering modes of objects in the 2.5D scene allows the user to investigate the information normally available only in a 3D representation of the scene (e.g. objects hidden around the corner).

The second textual based method focuses on presentation of graphical data with voice user interface. The possibility to investigate the structure of a 3D scene by means of textual queries where voice-based communication was used was successfully proven. A crucial role plays ontologies for storing data including natural language attributes and the usage of contextual information to improve the speech recognition rate and to resolve natural language ambiguities.

The hypothesis that the user's conversation language can be restricted accordingly to the conversation context determined by the application ontology and user context was proven. The dynamically generated input language of the voice user interface matched the user needs during the communication.

5.1. Future work

Our solution uses efficient information filtering based on the semantic description, which significantly increases the usability of the system in a mobile environment.

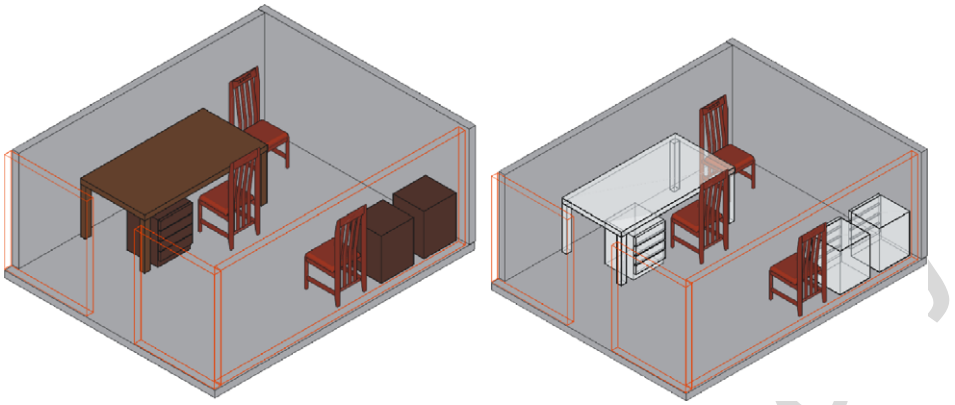


Fig. 19. Some of the furniture is described in semantic description as wooden chairs. The rendering mode of these chairs can be changed to highlight or suppress them.

One area of future work will investigate the possibility of utilization of the semantic description of the 3D scenes in the filtration process. The semantic description stored in the OWL [17] (Web Ontology Language) will be used for selective changes of the rendering modes of the objects in the 2D scene. The semantic description is usually application oriented—this means that the filtration process could be application driven. See an example in Fig. 19. This example shows the situation where only wooden chairs (application oriented information) are highlighted.

A second area of future work is finding a point with the optimal user experience with a compromise between the size of the input language and the recognition rate. One possible way is to introduce the behavior scenarios of the user in particular tasks (e.g. taking samples) and application domains (e.g. FM domain), which will help to predict the user actions and allow more precise restriction of the conversation language.

The goal is to develop a voice-based user interface, which will be usable in the real environment of a specific class of applications. For this purpose we plan to perform a second set of usability tests that would simulate the real work of a construction site inspector to find out whether input language being restricted in time really suits the needs of a real user.

Acknowledgements

The research was conducted within the framework of the MUMMY project (Mobile knowledge management—using multimedia-rich portals for context-aware information processing with pocket-sized computers in facility management and at construction site) and is funded by Information Society DG of European Commission (IST-2001-37365). See http://www.mummy-project.org/.

This research has been partially supported by MSMT under research program MSM 6840770014.

We appreciate very much the support provided by IBM research center Czech Republic for helping us to solve the problems of speech recognition and synthesis.

References

[1] Slavik P, Cmolik L, Mikovec Z. Object based manipulation with 3D scenes in mobile environment. Dagstuhl seminar proceedings 05181 mobile computing and ambient intelligence: the challenge of multimedia 〈http://drops.dagstuhl.de/opus/volltexte/2005/373/〉, 2005.

[2] Kopsa J, Mikovec Z, Slavik P. Ontology driven voice based interaction in mobile environment. In: Dagstuhl seminar proceedings 05181 mobile computing and ambient intelligence: the challenge of multimedia 〈http://drops.dagstuhl.de/opus/volltexte/2005/376/〉, 2005.

[3] Vainio T, Kotala O. Developing 3D information systems for mobile users: some usability issues. In: Proceedings of second NordiCHI'02 conference on human–computer interaction. New York, USA: ACM Press;2003. p. 231–4.

[4] Sanna A, Zunino C, Lamberti F. A distributed architecture for searching, retrieving and visualizing complex 3D models on personal digital assistants. International Journal of Human-Computer Studies 2004;60(5–6):701–16.

[5] Zunino C, Lamberti F, Sanna A. A 3D multiresolution rendering engine for PDA devices. In: Proceedings of seventh world multiconference on systemics, cybernetics and informatics (SCI03), vol. 5, 2003. p. 538–42 [ISBN9806560019].

[6] Hekmatzada D, Meseth J, Klein R. Non-photorealistic rendering of complex 3D models on mobile devices. In: Proceedings of eighth annual conference of the international association for mathematical geology, vol. 2, 2002. p. 93–8.

[7] Diepstraten J, Gorke M, Ertl T. Remote line rendering for mobile devices. In: CGI '04: Proceedings of the computer graphics international. Washington, USA: IEEE Computer Society; 2004. p. 454–61.

[8] Scalable vector graphics (SVG) 1.1 specification 〈http://www.w3.org/TR/2003/REC-SVG11-20030114/〉.

[9] Newell ME, Newell RG, Sancha TL. A solution to the hidden surface problem. In: ACM'72: Proceedings of the ACM annual conference. New York, USA: ACM Press; 1972. p. 443–50.

[10] Virtual reality modelling language (VRML) Specification 〈http://www.web3d.org〉.

[11] Ramakrishnan IV, Stent A, Yang G. HearSay: eabling audio browsing on hypertext content. ACM WWW2004, 2004. p. 80–9.

[12] Zue V. JUPITER: a telephone-based conversational interface for weather information. IEEE Transactions on Speech and Audio Processing 2000;8(1):100–12.

[13] Leong LH, Kobayashi S, Koshizuka N, Sakamura K. CASIS: a context-aware speech interface system. IUI '05: Proceedings of the 10th international conference on intelligent user interfaces. New York, USA: ACM Press; 2005. p. 231–8.

[14] Oviatt S. Advances in robust multimodal interface design. IEEE computer graphics and applications, vol. 23, no. 5. Los Alamitos, USA: IEEE Computer Society Press; 2003. p. 62–8.
[15] Pfleger N. Context based multimodal fusion. ICMI '04: Proceedings of the sixth international conference on multimodal interfaces. New York, USA: ACM Press; 2004. p. 265–72.
[16] Reithinger N, Alexandersson J, Becker T, Blocher A, Engel R, Löckelt M, et al. SmartKom—adaptive and flexible multimodal access to multiple applications. ACM ICMI'03. New York, USA: ACM Press; 2003. p. 101–8.
[17] Web ontology language (OWL) specification 〈http://www.w3.org/2004/OWL/〉.
[18] Kopsa J. Voice user interface for multimodal data. Master thesis, CTU in Prague, Prague, 2005.
[19] Cmolik L. Transformation of 3D scenes to 2D for mobile environment. Master thesis, CTU in Prague, Prague, 2005.

Appendix B

An evaluation tool for research of user behavior in a realistic mobile environment

Maly I., Mikovec Z., Vystrcil J., Franc J., Slavik P.: An evaluation tool for research of user behavior in a realistic mobile environment. In Int. Journal of Personal and Ubiquitous Computing. 2013, vol. 17, no. 1, p. 3-14. ISSN 1617-4909. **IF=1.113**

Pers Ubiquit Comput (2013) 17:3–14
DOI 10.1007/s00779-011-0475-0

ORIGINAL ARTICLE

An evaluation tool for research of user behavior in a realistic mobile environment

Ivo Maly · Zdenek Mikovec · Jan Vystrcil · Jakub Franc · Pavel Slavik

Received: 19 February 2011 / Accepted: 30 August 2011 / Published online: 14 October 2011

Abstract User behavior is significantly influenced by the surrounding environment. Especially complex and dynamically changing environments (like mobile environment) are represented by a wide variety of extraneous variables, which influence the user behavior in an unpredictable and mostly uncontrolled way. For researchers, it is challenging to measure and analyze the user behavior in such environments. We introduce a complex tool—the IVE tool—which provides a unique way of context visualization and synchronization of measured data of various kinds. Thanks to this tool it is possible to efficiently evaluate data acquired during complex usability tests in a mobile environment. The functionality of this tool is demonstrated on the use case "Navigation of visually impaired users in the building with support of a navigation system called NaviTerier." During the experiment, we focused on collection and analysis of data that may show user stress and which may influence his/her ability to navigate. We analyzed objective data like Galvanic Skin Response parameter (GSR), Heart Rate Variability parameters (HRV) and audio video recordings and also subjective data like the user's subjective stress feeling and observation of the user's behavior.

I. Maly (✉) · Z. Mikovec · J. Vystrcil · J. Franc · P. Slavik
Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
e-mail: malyi1@fel.cvut.cz

Z. Mikovec
e-mail: xmikovec@fel.cvut.cz

J. Vystrcil
e-mail: vystrjan@fel.cvut.cz

J. Franc
e-mail: francjak@fel.cvut.cz

P. Slavik
e-mail: slavik@fel.cvut.cz

Keywords User behavior · Context sensitivity · Measuring usability · A11y · User experience

1 Introduction

The research of user behavior in mobile environments faces a problem with the trade-off between the ability to measure all parameters in appropriate detail (possible mostly in laboratory environment only) and the ecological validity of the experiment (which can be ensured by field studies). Field studies performed in natural mobile environments (building, street, etc.) introduce two main problems. First, the measurement of needed behavioral parameters is rather difficult if not impossible in comparison with the laboratory environment; e.g., recording finger movement on the display or recording of the whole environment influencing directly the user behavior. Second, the complexity and dynamics of the natural environment with a wide variety of extraneous variables influence the user behavior in an unpredictable and mostly uncontrolled way.

On the one hand, we as researchers want to evoke this exact situation, which is ecologically valid and has potential to show us realistic user behavior in the mobile environment. On the other hand, there is a problem with the interpretation of such observed behavior as we cannot control all variables of the study.

The question is whether we are able to measure a sufficient amount of data with sufficient precision and whether we are able to analyze these data in such a way that we could correctly interpret the behavior observed.

In this paper, we will introduce a complex tool—the IVE tool—for the visualization of multiple data sources (generated by various measurements of user behavior) and evaluation of participant behavior in a natural mobile

environment. The benefit of unique context visualization plug-in and synchronization functionality during the evaluation process will be demonstrated by the use case "Navigation of visually impaired person inside buildings" with support of a navigation system called NaviTerier. This paper demonstrates efficient evaluation of data acquired by a complex usability test in the mobile environment by means of a special evaluation tool called IVE.

2 Use case: navigation of visually impaired person inside buildings

The use case we have chosen to demonstrate the analytical IVE tool and the process of evaluation and interpretation of user behavior observed in the mobile environment is "Navigation of visually impaired person inside buildings." The visually impaired person is walking through the building from his/her actual position (for example, the entrance) to the requested destination with help of the navigation system called NaviTerier. NaviTerier is a mobile interior navigation system for visually impaired users running on a standard mobile phone typically used by visually impaired persons. The main principle of the system is based on well-prepared descriptions of the route in the interior suited to the needs of visually impaired users. The NaviTerier application is utilizing standard Text To Speech application installed on the user's cell phone. A description of the complete route is divided into logical segments that are reproduced step by step to the user as he/she is moving through the route (see Fig. 1).

3 Early stages of research

The main research issue, besides the usability of the system control and understandability of the navigation description, was the question: What are the most important orientation cues that should be presented to the visually impaired user to ensure reliable and efficient navigation and orientation in an unknown area? The ideal way to answer this research question is to perform a usability study in a real environment. The task should be navigation to some destination in a building. After the walkthrough, the participants should be asked about the importance of the orientation cues on the route.

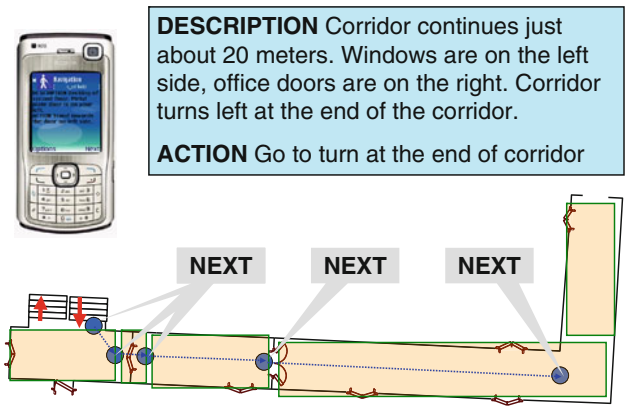


Fig. 1 NaviTerier—navigation principle

We have performed two qualitative studies to evaluate usability of control of our navigating system and also to improve descriptions of the environment to be valuable and easily understandable for the users. To analyze the importance of features (objects, shapes of the corridors, etc.) on the route and indicate the candidates for important orientation cues, we have also performed post-test interviews to get feedback from participants. We have asked the participants to try to recall the whole route they have passed in order to check our assumption that the most important features will be remembered much better than the less important ones.

Surprisingly, we have observed that the ease of remembering the features on the route is dependent not only on the type of the feature but also on their position on the route. In particular, we have observed that there were parts of the route where the participants were not able to recall features (like doors, shape of the corridor) they were correctly reporting in the rest of the route. We have made an assumption that this situation could be caused by a stress reaction, which negatively influences the cognitive processes, especially attention and memory performance in its acquisition phase [1].

4 Validity issues

There is a whole range of techniques for studying spatial environmental knowledge and its acquisition both by sighted and visually impaired persons. Most popular are virtual (for example Foo et al. [13], Gillner and Mallot [19], Kjeldskov [6]) or micro-scale artificial environment settings (for example, Tellevik [20] or Ochaíta and Huertas [21]) that allow researchers to control the environment settings and parameters to a great extent. When studies are conducted in real environments, they mostly focus on short routes rather on complex real-world-like situations. Studying spatial knowledge acquisition in real-world large-scale complex environments is very rare. Such an approach brings numerous methodological challenges as researchers do not control the environment settings and parameters so well.

Kitchin, Jacobson [9] state that the wayfinding in large-scale real-world spaces is different from wayfinding in limited areas. Natural large-scale environments provide different sources of environmental cues and a dramatically

different amount of environmental information than the artificial limited settings. They also show distractions based on situational context (other people, obstacles, diversions along the route). Studies conducted in artificial and small-scale environments suffer from the lack of ecological validity that prevents the research findings to be extrapolated into wayfinding and spatial knowledge acquisition in the daily lives of the visually impaired [9].

Our research approach follows Jacobson's and Kitchin's call [9] to "start to assess the knowledge and abilities within complex real-world environments that everyone inhabits, rather than inferring that results from the laboratory will exist in natural settings." This approach will require the collection of much larger amounts of data of various natures that should be efficiently evaluated.

5 Experiment

In this section, we will describe an experiment, which will serve as a demonstration of problems that the researcher has to face when observing user behavior in a realistic mobile environment. We will define research topics and requirements on the measurement methods we need to use to gather appropriate data for evaluation.

5.1 Research topics

On the basis of the previous research, we have formulated the following research topics:

1. Is it true that the acute stress reaction following the stress stimuli influences negatively the ability to remember the features on the route?
2. Which features and to what extent are they affected by the decreased ability to remember caused by the stress reaction?

To investigate these topics, we needed to setup an experiment where we will be able to generate stress stimuli, check whether there will be invoked a stress reaction of the participant and to get information about the features remembered by the participant. The experiment setup had to calculate and to perform the test with two groups—an experimental and a control group. In the following sections, we will describe the measuring methods used, the test route preparation and the experiment procedure.

5.2 Measuring methods

5.2.1 Measuring stress

We have studied several approaches for stress measurement to find those which will accommodate the following criteria:

- Non-invasive
- Continuous logging of potential stress level
- Resistant to direct influence of physical activity
- Suitable for mobile setup

Common methods based on measuring of urinary excretion rates of noradrenaline, adrenaline or salivary cortisol levels [14] were not suitable for our purposes because of the necessity of continuous detection. Analysis of pupil diameter to detect stress is not relevant due to target group of visually impaired users [15].

The most suitable methods to satisfy the above-mentioned requirements appear to be the measurement of Galvanic Skin Response parameters (GSR) and Heart Rate Variability parameters (HRV).

5.2.2 GSR analysis

Analysis using GSR is a method based on measuring of skin conductance that is rapidly changing in dependence on physiological changes caused by stress [11]. Stress activates the sympathetic nervous system, thus resulting in increased levels of sweat in the sweat glands and consequently increases the electric conductance of the skin [2]. Our experiment route (see Sect. 5.3) was designed in a way that the stress stimuli are the only variables in the experiment, and we expect that the GSR values will be primarily influenced by these stimuli. Appearance of other emotions that could influence GSR was checked using other data sources like subjective evaluation and observation of audio/video recording.

In the experiment, we have used the GSR sensor BodyMedia SenseWear PRO[1] in the form of a band that is placed on the back part of the upper arm. Data are logged to internal memory of the sensor. It also measures body temperature, accelerations, etc.

5.2.3 HRV analysis

HRV analysis is a method based on ECG signal analysis, where specific changes in heart rate reflect physiological changes caused by stress [16, 17]. According to [2], HRV represents the variations in the beat-to-beat intervals. HRV analysis is prevalently used to assess the effect of autonomic regulation on the heart rate. It provides a dynamic nature of the interplay between the sympathetic and parasympathetic branches. The relation between the sympathetic and parasympathetic branches can be described by the ratio of low and high frequencies LF/HF[2] in frequency spectrum obtained from beat-to-beat intervals in time.

[1] http://sensewear.bodymedia.com.

[2] According to [10] LF/HF parameter of HRV, LF = 0.04–0.15 Hz and HF = 0.15–0.4 Hz.

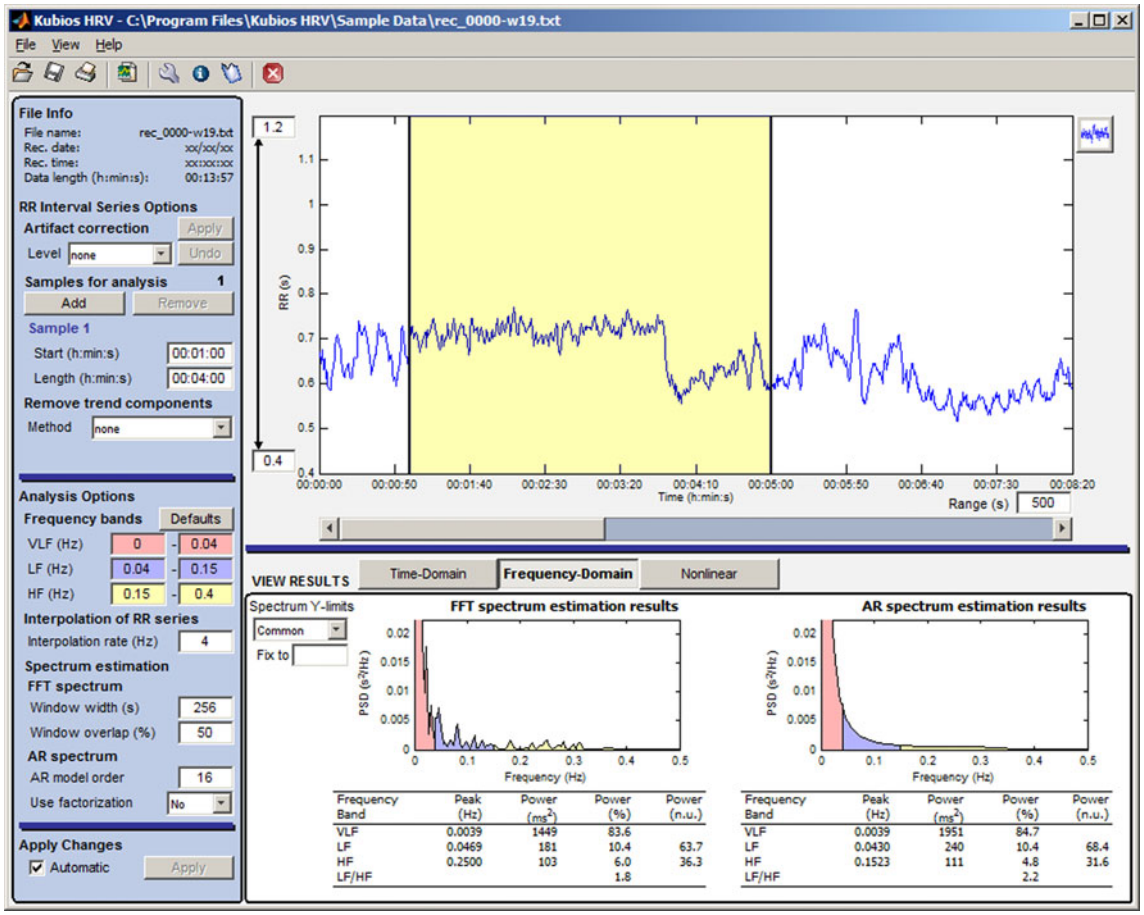


Fig. 2 Kubios HRV with 4 min sliding window

These frequencies are obtained by means of FFT (Fast Fourier Transformation). There is a balance between the sympathetic and parasympathetic branches under normal situations, placing the body in a state of homeostasis. However, under a state of mental stress, this balance will be altered. These states can be observed as differences between the values of LF/HF ratio in time. Each value (LF/HF ratio) we calculate is based on the analysis of certain interval of signal (sliding window). We have been working with a 4-min-long[3] sliding window to get insight into how LF/HF HRV parameters change in time. In the experiment, we have used Kubios's HRV [5] application to calculate LF/HF ratio parameters (see Fig. 2 for highlighted one 4-min window from 1:00 to 5:00).

We have also used an ECG sensor in the form of a chest strap[4] that was installed on the participant's body, and a wire connected logging unit has been placed in a small backpack carried by participants [18].

5.2.4 *Subjective evaluation*

During the post-test phase, the subjective feeling of the stress was evaluated. The route, through which the user was walking, has been divided into 6 parts that were described to participants by the evaluator. Participants were invited to sort them in descending order according to their subjective feeling of stress in those parts. The force choice technique (each segment had to be ordered) was used. The forced choice principle was used to identify the most stressful segment for later comparison with other data sources.

5.2.5 *Camera*

In order to record the exact movement of the participant, a small camera has been attached to a shoulder strap of the backpack. The camera recorded the area in front of the participant. Participants were also recorded by a third person for possible better analysis of the context of participant behavior.

[3] Bayevski [10] uses 5-min-long sliding window. As we have relatively short measurements of signal, we have used 4-min sliding window to gain more samples of LF/HF ratio.

[4] In fact, it was modified ECG sensor from Polar®.

5.2.6 Observation in the field

Physical interaction with all objects on the route was registered in the log sheet. These data were gathered for further comparison with the objects mentioned by the participant in the interview during the post-test phase.

5.2.7 Logging

The NaviTerier application is logging all interactions the user has performed with the mobile device so we do not need any direct screen capturing of the display.

5.2.8 Post-test interview

After completion of the test route, participants were asked to freely describe the route verbally to gain insight into how they remember the route they have taken. All remembered objects had to be mentioned and were registered in the log sheet.

5.3 Test route preparation

Our plan was to prepare a route identical for both groups of users (experimental and control). The difference would be in one segment only, which would contain real stress stimuli for the experimental group and no stress stimuli for the control group. All other segments had to be the same. Unfortunately, we were not able to find a route that would comply with these requirements. But according to Gray [12], we can simulate the stress environment by verbal and physical stimuli with a similar effect as in a real stress environment [12]. Therefore, we decided to force evocation of stress in one of the segments for the experimental group of participants, and all other parts of the route remained unchanged. The second group of participants was treated as the control group.

Forced evocation of stress has been performed through realistic auditory and haptic stimuli associated with a potentially threatening environment. In one of the corridors, reconstruction work has been simulated. Participants have been warned by the NaviTerier of potential appearance of dangerous obstacles. The floor was covered with plastic foil, and a vertical barrier from the plastic foil was installed across the corridor. Buckets and other instruments were installed on the floor to ensure a more realistic sense of reconstruction. The control group met only the standard corridor without any obstacles in the mentioned corridor.

5.3.1 Placing stressful stimuli

Expecting individually different reactions to the stress stimuli, we have chosen to measure the level of stress objectively through GSR and HRV methods as a verification mechanism.

Placing stressful stimuli in one of the segments of the route did not serve as a direct precursor of stress reaction. When initially analyzing both HRV data and subjective reports of the participants, we have found that the unsystematic and uncontrolled situational variables on the route (e.g., random social encounters, environmental features such as distant noises that could work as navigation cues) were causing significant stress levels in both experimental and control groups.

5.4 Test procedure

In the experiment, we were observing two groups of users. In the experimental group, we observed 12 blind participants, and in the control group, we observed 10 blind participants. The complete experiment consisted of 5 consequent phases (pre-test phase, training phase, break, test phase and post-test evaluation).

5.4.1 Pre-test phase

During the pre-test phase, participants were acquainted with the study procedure. The principle of the NaviTerier navigation system was explained, and instructions on how to operate a cell phone with the NaviTerier application were given to the participants. All sensors described in Sect. 5.2 (ECG, GSR, camera) were installed on the participant's body.

5.4.2 Training phase

During the training phase, participants went through a training route that was totally different from the test route. The training phase took approximately 10 min, and participants were followed by test conductors to help with any problems that might occur. The main purpose of the training phase was to get participants familiar with the navigation device and to avoid any technical problems during the test phase. Only the arm placed GSR sensor was turned on in this phase as it needs some time to "warm up" for correct measurement. After passing the training route, participants were led back to the laboratory to have a rest, settle down and to be instructed for the test phase.

The ECG measurement system and audio–video recording had been switched on. Participants were informed that they will go through the test route independently and also given important instructions to try to remember the route precisely so next time they could go through the route on their own.

5.4.3 Test

The whole test route consisted of 15 segments including stairways, corridors and doors (see Fig. 3). Participants were led at the start of the first segment that was a square-shaped stairway. It had to be stepped 2 floors down. To avoid any potential problems with opening of doors, all doors on the route were kept opened.

After passing the first doors, there was a so-called "stress segment" highlighted in Fig. 3. Depending on whether the participant was in the experimental or control group, there was either the setup as described in Sect. 5.3 or just a normal corridor. The rest of the route consisted of corridors, turns and doors.

To avoid any potential disruption of the test, there was special emphasis on carrying out all observations (third person audiovisual recording and field observation of participant interaction with the environment) without being noticed by the participants.

5.4.4 Post-test evaluation

Participants were returned to the laboratory, and all sensors were removed from their body. They were asked to describe the route verbally to gain the information about how they remember the route features they have passed. All remembered features (objects, corridors, crossings, etc.) had to be mentioned and were registered in the log sheet.

Six parts of the route (defined in Sect. 5.2) were described to participants. Participants were then invited to sort these parts in descending order according to their subjective feeling of stress in those parts. Forced selection principle had been used so no two parts could be marked with the same stress level.

Participants were informed that a video of their walk through had been taken by a third person and they were asked for permission for using these video recordings for further analysis.

6 Analysis and scientific evaluation of measurements

In order to perform scientific evaluation that is typically determined by research topics, we need tools that will allow us to perform efficient human-based observation and analysis of measured values. The analysis is typically a process, where we are trying to find interesting patterns and correlations in the data. The analysis could be facilitated by the ability to define combinations of various parameters and compare this newly derived information with other previously measured data.

This scientific evaluation process must be supported by complex visualization because it involves analysis of a wide range and large amount of data available to the researcher. This is additional to the data measured by the experimental setup also providing contextual data (like state of the NaviTerier application, environment surrounding the test participant), which are essential for the interpretation of observed behavior patterns. These data are typical of a very different nature and data types, e.g., logs, AV data, questionnaires (externalization and subjective stress), observation reports (incidence with objects on the route). The tool must be able to cope with the processing, visualization and synchronization of these data sources.

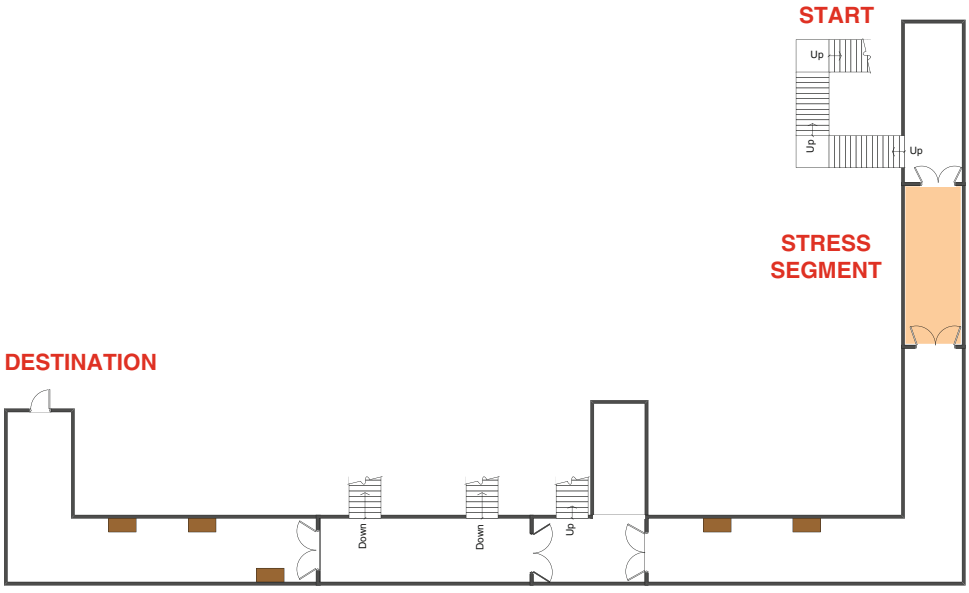


Fig. 3 Test route schema

For scientific evaluation, it is also essential to be able to inspect the visualized data from different points of view to search for interesting behavioral patterns.

Finally, it must be possible to analyze the data across several participants to support the analysis of similarities or differences in user behavior or in other collected data. We would like to see selected data side by side, with possibilities to align them or to highlight same situations, e.g., same tasks or similar stress data trends.

- The analytical tool requirements can be summarized as follows:
- the length of necessarily observed video must be efficiently minimized
- only video is viewed sequentially, other data should be visible at a glance with possibility to search, directly access any part of data and focus on details
- data views must be synchronized
- wide range of data of very different types must be visualized
- annotation functionality must allow efficient derivation of new information from existing one
- any data source must be able to serve as master data, where the focus of the researcher is paid and other data must adapt synchronously

From the data source point of view, it is necessary that the analytic tool can work with the following data sources:

- application model of the tested application
- log files from tested application
- observation of interactions with objects (e.g., doors, flowers, windows) during the test
- post-test analysis (e.g., subjective evaluation of stress level on the route, remembered objects the participant interacted with on the route)
- HRV data recorded during study and analyzed using Kubios HRV [5]
- GSR data and two audio/video recordings

We found out that most tools for visual data analysis offer only static visualization of generic spreadsheets or database data in the form of predefined visualizations, like plots, maps, graphs or dashboards. Some tools (e.g., VisiFire [7]) allow for animations of the data changes. Generally, none of these tools allow synchronous visualization of the spreadsheets or database data together with multimedia files, as in. video recordings. Observer XT tools from Noldus [8] allow complex analysis of data from behavioral studies together with multiple video recordings. However, the Observer XT is not able to provide us with custom visualizations of data like visualizations of application states in the form of segment view or show us non-standard temporal data (with no direct mapping of values to time line), like HRV values from Kubios, synchronized with video recording.

Because we have not found any suitable analytic tool, we decided to use the IVE (Integrated Interactive Information Visualization Environment) tool and develop new plug-ins for data source import and data visualization that will suit our test. The tool was in detail presented in [3]. An advantage of the IVE tool, compared to generic visualization toolkits like VTK, is that it allows quick creation of custom data importers and custom visualization plug-ins synchronized with multimedia files. It is designed for the analysis of temporal data, and therefore, it is easier to develop synchronized visualizations in it. Both the IVE tool and the plug-ins used and developed for the required data source visual analysis are described in the following sections.

6.1 Description of IVE tool structure

IVE is an interactive tool for visual analysis of data from usability studies. IVE tries to give the usability expert the power of the qualitative analysis tool to understand user behavior but it also allows statistics calculation.

The IVE tool is based on the structure of Information Visualization Reference Model [4]. This model represents a framework suitable for implementation of various information visualization applications. A problem with tools for visual analysis is that in each study the data sources are in a different format and/or different types of data are available. The main advantage of the IVE tool is that it successfully copes with both problems using plug-ins.

In Fig. 4, we can see how the process works. IVE uses an internal object database and converter plug-ins to convert data from raw sources into this database that basically contains records of key/value pair or multimedia data. Because generic objects in the internal database would bring unnecessary complexity to both the conversion process and reusability of visualizations, we limit the type of internal data types. Currently, the IVE works with:

- audio/video recording,
- generic log file (list of records, each record contains list of key/value pairs),
- task model,
- application model.

Depending on the structure of available data, the usability expert can use a subset of visualization plug-ins or a subset of the visualization plug-in functionality. This means that no data source is mandatory, but when more data sources are available, the IVE is able to take advantage of this fact and can allow interconnection of data and therefore enhancement of visualizations, e.g., interconnection of navigation segment visualization with

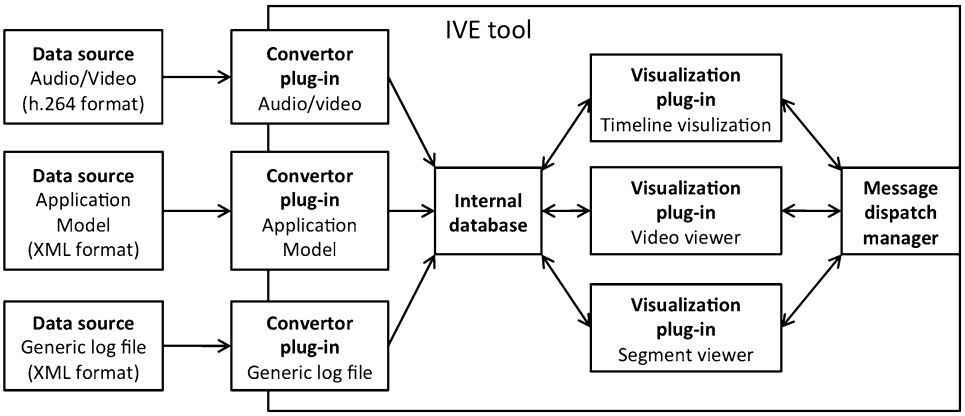


Fig. 4 Schematic structure of the IVE tool and IVE convertor and visualization plug-ins

navigation segment instructions for test participant and subjective stress rating of that navigation segment.

Because IVE limits the type of data stored in its internal database, it is easier to develop new visualization plug-ins for IVE. Each visualization plug-in is developed as a plug-in with one or more views that have access to the IVE internal database, and it can communicate with other plug-ins through a simple message dispatching system. The IVE tool with visualization plug-in views developed for the analysis of data from NaviTerier navigation application is shown in Fig. 5. Details of the plug-ins functionality and data visualization are in the following sections.

A visualization plug-in is instantiated using the IVE wizard that allows selection of the visualization plug-in and selection of data sets (log file, application model, audio/video recording) that will be used in the visualization plug-in views. Each visualization plug-in can announce which data sets are mandatory, and it can also limit the amount of data sets. Some visualization plug-ins may announce changes to other visualization plug-ins, e.g., highlighting of the same information in other visualization plug-in views. Instead of storing this information in the internal database, we used a central message dispatcher that collects and resends all messages between all visualization plug-ins.

The IVE tool was developed in Java using the NetBeans Platform, and therefore, it includes several features typical for this platform. First of all is the use of the built-in plug-in update manager, which is used for all plug-ins for the IVE tool and which allows for unified handling of plug-ins. Second, the tool uses a window management of the NetBeans platform that allows for a complex placement and manipulation or even undocking of the visualization plug-in views in the main panel of the tool.

6.2 Timeline visualization plug-in

The Timeline visualization plug-in is the plug-in that shows a combination of several data sources in one view (see Fig. 6). This plug-in is able to show the log file of the application (NT Segments timeline), additional user

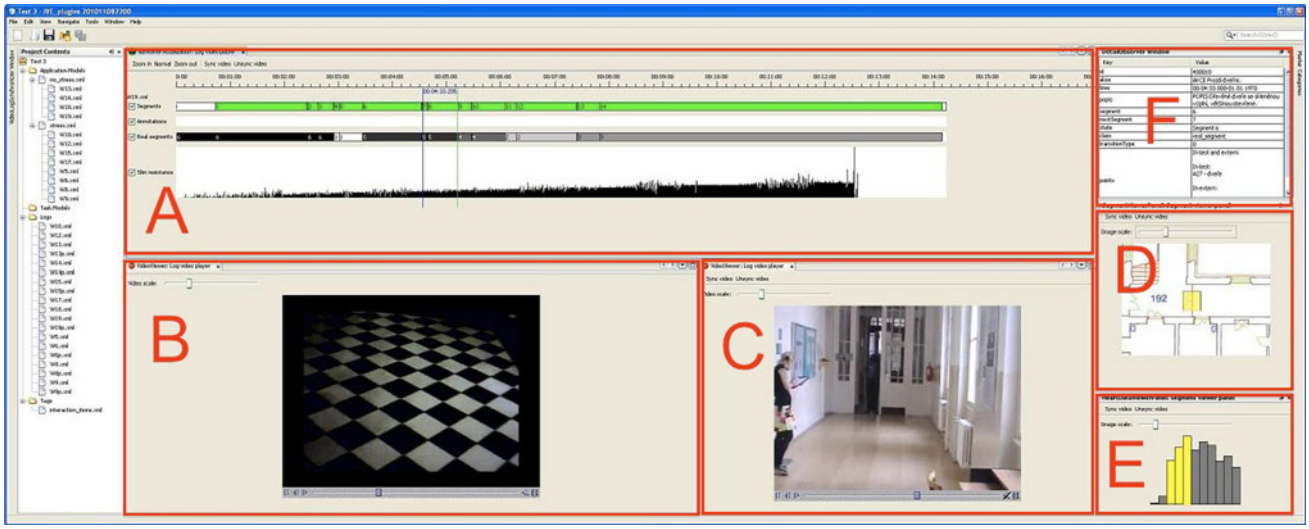


Fig. 5 IVE tool with Timeline visualization plug-in (*a*), two video viewer plug-ins (*b*, *c*) Segment viewer plug-in (*d*), HRV viewer plug-in (*e*) and Detail window (*f*)

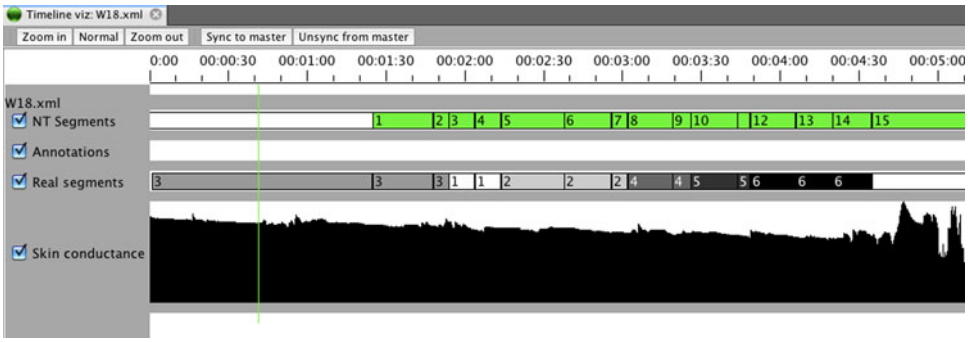


Fig. 6 Timeline visualization

annotations (Annotations timeline), the real position of the user in a segment (Real segments timeline) and GSR data (Skin conductance timeline). For each record in each line, there is the possibility to show details in the Detail window (described in following sections). Each timeline can be hidden with a checkbox in front of the timeline name, zoomed in and out using buttons on the tool bar and synchronized with the master video player. When the timeline is synchronized there is an indicator for the position in the video and current log file details are presented in the Detail window.

Each timeline uses a different visualization to present data. The NT Segments timeline shows rectangles that represent selection of new segments in the tested application. The color represents the type of action. e.g., green color (or ascending sequence of following rectangle labels) means that the user selected the next segment and he/she is probably following the test smoothly. Red color (or descending sequence of following rectangle labels) means that the user had to select the previous segment. That may signal that the user had problems and such a point should be further analyzed in the video.

The Real segments timeline shows the real position of the user in the segment. This position may differ from application segments when the user is lost or when he/she starts the next segment during his/her way to the next segment. The hue and the label represent the value of the subjective stress rating of the particular segment, which was collected after the test. White color (1) is lowest stress; black color (6) is highest stress.

The Skin conductance timeline shows values of the GSR, which are scaled into the height of the line view. The exact values are shown in the Detail window (described in the following section).

The Annotations timeline shows additional annotations added during the analysis of data in the IVE tool. For each annotation, there is a rectangle that represents start and end time of the annotation. Annotation details are shown in the Detail window. Each annotation (see detail of annotation editor in Fig. 7) can store annotation description and a set of tags that may be used for filtering of the annotations.

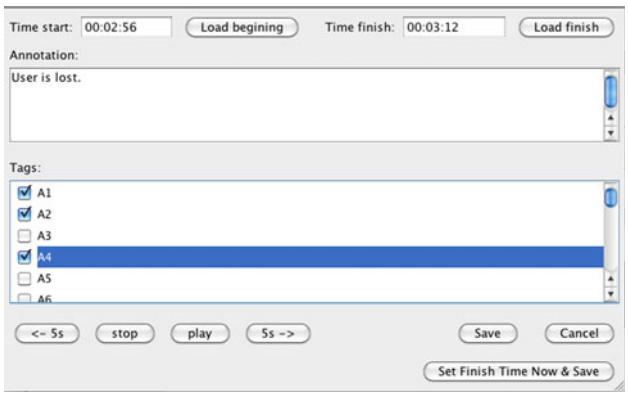


Fig. 7 Annotation window detail

There are also basic video controls for easier setting of start and end time and for repeating observation of a particular point.

Apart from the data analysis, the Timeline visualization plug-in is used also for time synchronization of timeline data sets with video. All data sets in the Timeline visualization plug-in use time of the day so we can easily transform this time into the time of the master video. For this purpose, we use the Timeline video synchronization plug-in (see Fig. 8). We find the same time point in the Timeline visualization plug-in view and in the Video plug-in and add it to the Video log synchronizer window and synchronize.

6.3 Multiple video visualization

The video viewer plug-ins are the plug-ins for sequential replay of all video files. One video viewer plug-in always acts as a master synchronization point, i.e., all other plug-ins synchronize to the master time. In case we have more then one video source, we can use other video viewer plug-ins that are then also synchronized to the master video viewer plug-in. The video viewer plug-in contains a scale slider, which allows adjustment of the video depending on the location in the IVE tool. The Slave video viewer plug-in also contains buttons for synchronization with the master

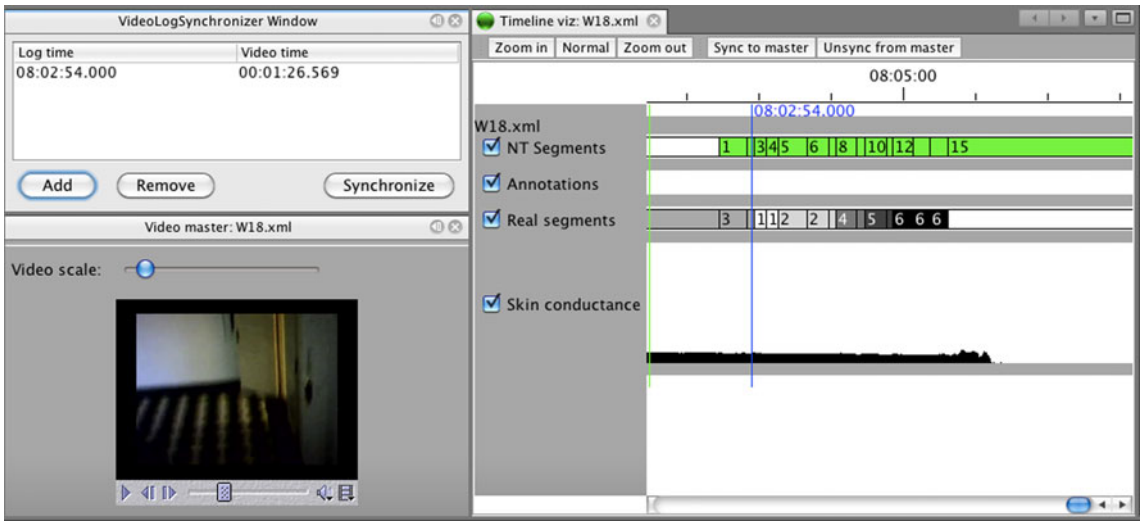


Fig. 8 Video log synchronization plug-in

Fig. 9 Slave video viewer plug-in

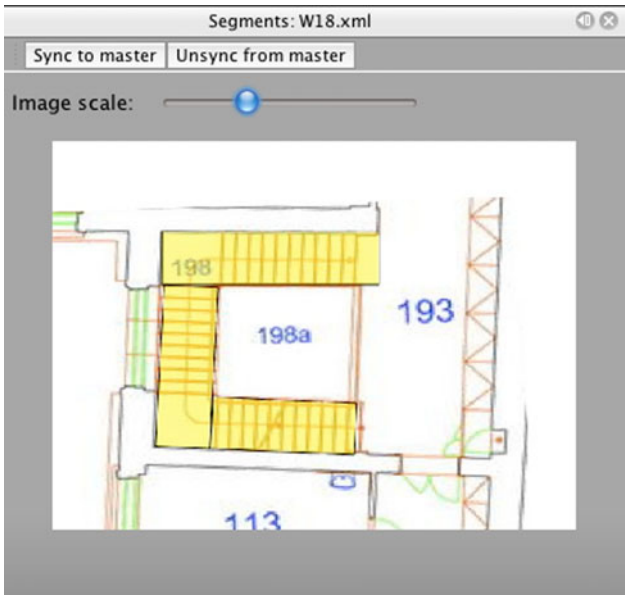


Fig. 10 Route segment visualization plug-in view

video viewer plug-in (see Fig. 9). During the test's data analysis, we used one master view. The number of slave views is limited by the resources of the PC and operating system.

6.4 Segment visualization plug-in

Segments are visualized mainly in the timeline. However, there is additional visualization of the route segment area in the form of image (see Fig. 10). The segment visualization is synchronized with the master video and with the real segment data source to show the current segment. There is also a scale slider, which allows for adjustment of the plug-in view depending on the location in the IVE tool.

6.5 Heart Rate Variability parameter visualization

As the HRV parameter analysis is based on the FFT, the obtained values are not assigned to a certain time. In the IVE tool, the Heart Rate Variability (HRV) visualization plug-in shows HRV data sets in the form of a bar graph (see Fig. 11). The first bar represents the LF/HF ratio for the first 4 min (4 min sliding window of FFT) of the analyzed signal. The second bar represents the LF/HF ratio for signal from time 1:00 to 5:00 min. Every other bar represents sliding the window by 1 min.

Exact LF/HF ratio value is shown in Detail window. The color of the bar is yellow when the actual time of video

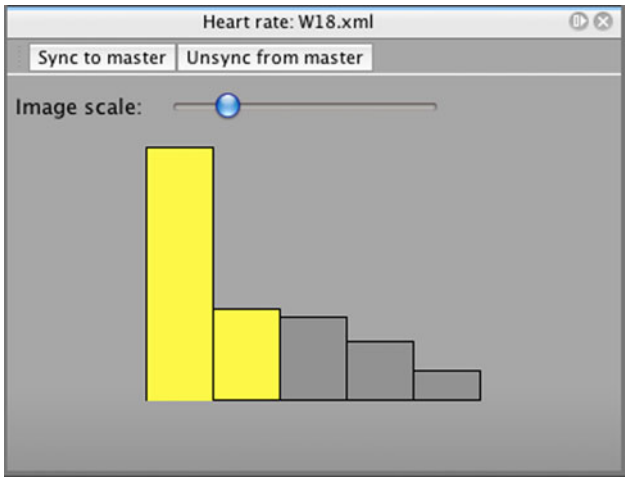


Fig. 11 HRV data visualization plug-in view

playback in the Video viewer plug-in intersects the corresponding sliding window interval. Therefore, up to 4 bars can be indicated as "active" and highlighted in yellow.

6.6 Detail window

The Detail window is a multipurpose window for visualization of plug-in record details. In the window, we show a list of key/value pairs that are sent by the particular visualization plug-in. In Fig. 12, you can see detail information for the segment. There are navigation instructions derived from the application model data set. There are also interaction points (points) that the participant interacted with during the test and that the participant mentioned in the after-test interview.

6.7 Performed analysis

In the test, we collected data from 22 participants. The length of the tests was between 8 and 30 min. For each participant, we got a log file from the application, his/her GSR data (from approx. 6,000–19,000 records), HRV data (from 4 to 20 sliding window values), log sheets from the test and from the post-test route description. For each participant, we recorded 2 videos.

The time for data synchronization was a maximum of 5 min, and the speed of annotation corresponded with the findings from our previous studies with the IVE tool [3].

7 Conclusion

In this paper, we have demonstrated tools and methods for efficient evaluation of experiments in mobile and realistic environments. Our demonstration use case was the "Navigation of visually impaired person inside buildings."

DetailObserver Window

Key	Value
segment	4
popis	DESCRIPTION Corridor, about 8 meters long. On the right there is door to toilete. The corridor turns right at the end. Before the end of the corridor, there is extinguisher on the right.
nextSegment	5
lod	0
state	Segment 4
class	real_segment
transitionType	0
id	400008
akce	ACTION Go to the end of the corridor and turn right.
time	08:03:16.000-01.01.1970
edge_id	143
command	Dalsi
points	In test and extern: In test: In extern: Not metioned: A8 - door (WC) A9 - cabinet A10 - embrasure A11 - extinguisher

Fig. 12 Detail window

During such experiments, a much larger amount of data of various kinds is generated in comparison with static and laboratory-based experiments. We have shown that for these kinds of experiments other methods for processing, visualization and analysis should be used. More specific, our analytical tool IVE is capable of handling huge amounts of data of very different types. It offers advanced features for synchronization of data (e.g., synchronization of audio/visual data with HRV data that does not have exact time assignment), visualization and analysis of contextual data like application state, graphical representation of the participant's surrounding and interaction of the participant with objects on the route.

Thanks to the IVE tool we were able to analyze the stress measurement data in the context of uncontrolled situational variables that potentially cause stress and distinguished them from a true indication of stress. These data were omitted from further analysis. Without the annotation plug-ins and advanced synchronization features of the IVE tool, we would be unable to detect the false indicators.

Through the global view on the various data provided by the IVE tool, we were also able to detect other stress stimuli that were hidden to us from simple watching of videos (e.g., GSR values were higher when the user had to

concentrate enormously on the NaviTerier route description).

During the experiment preparation and evaluation, we found that we need some very special plug-ins (e.g., visualization of participant surroundings during route walkthrough). Thus, we appreciated the efficient plug-in interface of the IVE tool, which in less than 3 weeks allowed us to develop 5 new plug-ins (described in Sect. 6).

Although the IVE tool was evaluated as useful, there are still some areas where we want to focus in future work. We found that applications with a complex layout of windows, like IVE, are quite difficult to set up and there should be assistance in the setup and saving of the window layout. Also, there should be easier switching between the data set visualization, e.g., switching between 2 users. Currently, we have to reload all visualization plug-in views. From the test point of view, we want to focus on analysis of interactions between the user and other objects during the test, e.g., visualize such interaction in a segment view plug-in.

Acknowledgments This research has been partially supported by the MSMT under the research program MSM 6840770014. This research has also been partially supported by the MSMT under the research program LC-06008 (Center for Computer Graphics). We would like to thank P. Smrcka and R. Kliment from the Joint Department of Biomedical Engineering CTU and Charles University in Prague, Studnickova 7/2028 Praha 2 for consultation and lending of the ECG for measuring HW and SW. We would like to thank the BioDat research group, Gerstner Laboratory, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague for lending of the GSR measuring device.

References

1. Sandi C, Pinelo-Nava MT (2007) Stress and memory: behavioral effects and neurobiological mechanisms. Neural Plast 2007: 78970. doi:10.1155/2007/78970
2. Salahuddin L, Desok K (2011) Detection of acute stress by heart rate variability using a prototype mobile ECG sensor. http://vega.icu.ac.kr/~kimdesok/IEEE-CS_IEEE-251.pdf. Accessed on 12 Feb 2011
3. Maly I, Mikovec Z, Vystrcil J (2010) Interactive analytical tool for usability analysis of mobile indoor navigation application. In: Proceedings of the 3rd international conference on human system interaction (HSI 2010), pp 259–266, doi: 10.1109/HSI.2010.5514559
4. Card SK, Mackinlay JD, Shneiderman B (eds) (1999) Readings in information visualization: using vision to think. Morgan Kaufmann Publishers Inc., San Francisco, pp 686. http://dl.acm.org/citation.cfm?id=300679 or http://www.amazon.com/Readings-Information-Visualization-Interactive-Technologies/dp/1558605339
5. Kubios HRV (2011) University of Eastern Finland, http://kubios.uku.fi/, Biosignal Analysis and Medical Imaging Group. Accessed 10 April 2010
6. Kjeldskov J, Skov MB, Als BS et al (2004) Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In: Proceedings of the MobileHCI conference 2004, pp 61–73. doi:10.1007/978-3-540-28637-0_6
7. Visifire (2011) http://www.visifire.com/. Accessed 1 Feb 2011
8. Noldus (2011) Human behavior research. http://www.noldus.com/human-behavior-research. Accessed 26 Jan 2011
9. Kitchin R, Jacobson RD (1997) Techniques to collect and analyze the cognitive map knowledge of persons with visual impairment or blindness: Issues of validity. J Vis Impair Blind 91:360–376
10. Bayevsky RM, Ivanova GG, Chireykin LV, Gavrilushkin AP, Dovgalevsky PYa, Kukushkin UA, Mironova TF, Priluzkiy DA, Semenov UN, Fedorov VF, Fleishmann AN, Medvedev MM (2002) HRV Analysis under the usage of different electrocardiography systems (methodical recommendations). These methodical recommendations are prepared according to the order of the Committee of Clinic Diagnostic Apparatus and the Committee of New Medical Techniques of Ministry of Health of Russia (protocol 4 from the 11th of April, 2002), Moskva http://www.vestart.ru/atts/1267/24baevsky.pdf. Accessed 1 Feb 2011
11. Benoit A, Bonnaud L (2009) Multimodal focus attention and stress detection and feedback in an augmented driver simulator. Pers Ubiquitous Comput 13(1):33–41. doi:10.1007/s00779-007-0173-0
12. Gray JA (1987) The psychology of fear and stress. Cambridge University Press, Cambridge
13. Foo P, Warren WH, Duchon A, Tarr MJ (2005) Do humans integrate routes into a cognitive map? map versus landmark-based navigation of novel shortcuts. J Exper Psych Learn Memory Cognition 31:195–215
14. Bassett JR, Marshall PM, Spillane R (1987) The physiological measurement of acute stress (public speaking) in bank employees. Int J Psychophysiol 5(4):265–273. doi:10.1016/0167-8760(87)90058-4
15. Yamanaka K, Kawakami M (2011) Convenient evaluation of mental stress with pupil diameter. http://www.ciop.pl/33712 Accessed 1 Feb 2011
16. Ohsuga M, Shimono F, Genno H (2001) Assessment of physical work stress using autonomic indices. Int J Psychophysiol 40:211–220. doi:10.1016/S0167-8760(00)00189-6
17. Berntson GG, Bigger JT, Eckberg DL et al (1997) Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology 34(6):623–648. doi:10.1111/j.1469-8986.1997.tb02140.x
18. Hána K, Smrčka P, Kašpar J, Fiala R, Mužík J et al (2009) The system for monitoring of the human psychophysiological state utility model No. 19422, Czech Industrial Property Office
19. Gillner S, Mallot HA (1998) Navigation and acquisition of spatial knowledge in a virtual maze. J Cogn Neurosci 10:445–463. doi:10.1162/089892998562861
20. Tellevik JM (1992) Influence of spatial exploration patterns on cognitive mapping by blindfolded sighted persons. J Vis Impair Blind 86:221–224
21. Ochaíta E, Huertas JA (1993) Spatial representation by persons who are blind: a study of the effects of learning and development. J Vis Impair Blind 87:37–41

Appendix C

Collaborative Navigation of Visually Impaired

Balata J., Franc J., Mikovec Z., Slavik P.: Collaborative Navigation of Visually Impaired. In Int. Journal on Multimodal User Interfaces. 2014, vol. 8, no. 2, p. 175-185. ISSN 1783-7677. **IF=0.833**

J Multimodal User Interfaces (2014) 8:175–185
DOI 10.1007/s12193-013-0137-9

ORIGINAL PAPER

Collaborative navigation of visually impaired

Jan Balata · Jakub Franc · Zdenek Mikovec · Pavel Slavik

Received: 4 July 2013 / Accepted: 28 November 2013 / Published online: 14 December 2013

Abstract A navigation system for visually impaired users can be much more efficient if it is based on collaboration among visually impaired persons and on utilising distributed knowledge about the environment in which the navigation task takes place. To design a new system of this kind, it is necessary to make a study of communication among visually impaired users while navigating in a given environment and on their regularly walked routes. A qualitative study was conducted to gain insight into the issue of communication among visually impaired persons while they are navigating in an unknown environment, and our hypotheses were validated by a quantitative study with a sample of 54 visually impaired respondents. A qualitative study was conducted with 20 visually impaired participants aimed at investigating regularly walked routes used by visually impaired persons. The results show that most visually impaired users already collaborate on navigation, and consider an environment description from other visually impaired persons to be adequate for safe and efficient navigation. It seems that the proposed collaborative navigation system is based on the natural behaviour of visually impaired persons. In addition, it has been shown that a network of regularly walked routes can significantly expand the urban area in which visually impaired persons are able to navigate safely and efficiently.

Keywords Collaboration · Communication · Regular routes · Navigation · Visually impaired

J. Balata (✉) · J. Franc · Z. Mikovec · P. Slavik
Department of Computer Graphics and Interaction Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
e-mail: balatjan@fel.cvut.cz

1 Introduction

The main difference between navigation systems specially designed for sighted users and for visually impaired users lies in the level of detail of the environment description, and in the representation of the instructions.

Sighted users and drivers are bound to streets and roads, and the details contained in a description of these elements are sufficient for them. However, visually impaired users make use of different navigation features in the environment, and a description provided by a navigation system specially designed for sighted persons is not adequate for them [23].

Several options for navigation applications on smartphones are nowadays used by visually impaired persons, although they were originally designed for sighted pedestrians and/or for drivers. The most common is the Nokia Maps application, pre-installed on Symbian [21] smartphones, which is widely used by visually impaired users because of their hardware keyboard and their good screen-reader [8]. Other options are built-in navigation applications either in Android smartphones (version 4.0 and above) [1] with Explore-by-touch support for eyes-free interaction, or in iOS devices with VoiceOver [2].

Situations in which orientation is lost are mentally demanding, especially for visually impaired users, and it is necessary to analyse ways to help them by means of navigational aids. Some partial solutions to this problem are available in the form of:

- navigation call centres (e.g. Navigation Center of Czech Blind United—SONS [20] where navigation instructors and operators directly navigate visually impaired persons),
- voice-enabled navigation applications for smart phones [6, 14],
- custom hardware or wearable computer aids [13,19,26].

Fig. 1 Alice is a visually impaired person who has decided to visit a new art exhibition for the visually impaired in the city centre on a Saturday evening, but does not know the destination or the route well

The last of these three solutions can involve a wide range of sensors, e.g. cameras, headsets for communication, GPS [12], or even Microsoft Kinect [15], to analyse images and provide information for visually impaired users about their surroundings.

Although all these navigation systems are designed for visually impaired persons, none of them is ideal. Important considerations are that many custom hardware or wearable computer aids are in the prototype phase of development, and only a small proportion of all aids can resolve situations in which a visually impaired user gets lost.

The goal of our research is to explore ways to use distributed knowledge of an environment among visually impaired users, and to design a collaborative navigation system based on communication among visually impaired users.

1.1 Use case

Let us imagine that Alice is a visually impaired person who has decided to visit a new art exhibition for the visually impaired in the city centre on a Saturday evening (see Fig. 1). She knows the address, but she does not know the destination or the route well. She does not have a computer, and she uses a smartphone equipped with a GPS module and the installed screen reader for all her communication (calls, messages, emails, web). If she feels unsure about the route, or if she gets lost, she might be able to use several available navigation methods:

- ask strangers,
- use the voice-enabled navigation application on her smartphone,
- call the navigation centre for the visually impaired,
- get in touch with a navigation instructor,
- a call visually impaired friend who may know the destination,
- call an unknown visually impaired person with useful knowledge of the desired destination, who walks past the gallery every day on the route from his/her home to a nearby public transport stop to get to work (there may be more than one visually impaired person able to help in this way).

The navigation centre unfortunately does not have sufficiently good information to describe the way to the gallery, but it can at least provide information about public transport. It is too late to get help in advance from a navigation instructor (Alice decided to go to the exhibition at the last minute), and Alice does not know any friends who can give her a description of the route from the public transport stop to the gallery. In addition, Alice does not feel comfortable asking strangers (see Sect. 2, Fig. 7). A further problem is that the GPS signal received in that area is not accurate enough for precise navigation via smartphone. The last option for Alice is some unknown visually impaired person.

Identifying and selecting a visually impaired person who can provide a sufficiently high-quality description of the destination environment is not a trivial problem. Several parameters need to be taken into account:

- frequency of visits to the destination, date of the last visit to the destination, direction of motion, etc.;
- preferences of the visually impaired helper, e.g. the time schedule when he/she is available, maximum number of incoming requests, etc.;
- the preferences of the person who is asking for help, e.g. duration, category [25], onset of the impairment (early or late blind), and whether he/she uses a guide dog.

1.2 Problem description

A common problem for all the existing navigation solutions is the lack of an environment description especially created for visually impaired users, focusing on special navigation points and orientation cues [11,22,23], which visually impaired users need for safe and efficient navigation. A specific description of this kind is hard to obtain, as only trained navigation instructors can provide it in a fully satisfactory form. Unfortunately, there are typically only a very limited number of such instructors.

A description with obvious limitations due to the level of impairment [3,18,24] can also be provided by visually impaired persons. Typical navigation points and orientation cues (landmarks) provided by visually impaired persons themselves are:

- leading lines formed by edges of the pavement, handrails, corners of buildings, the better side of the street to travel on, etc.,
- sounds from traffic, construction work, water, width of the street from the echo, etc.,
- smells of various types, e.g. bakeries, drugstores, sewers, etc.,

– the direction of heat from the sun at certain times of day,
– approximate distances, changes in elevation, type of ground material.

For the names of streets, the order of turns on the route, navigation through parks and estates, the location of a specific house in the street, etc. visually impaired people need a description from a navigation instructor.

Visually impaired persons remember the special description of frequent routes provided by navigation instructors, and can pass these descriptions on to other people. There are many more visually impaired users with specific knowledge than there are instructors available. The problem is that their knowledge is limited to a few routes, and they cannot generate new complete navigation descriptions without the help of an instructor.

An idea for overcoming these problems is to support and facilitate direct simultaneous help between visually impaired people, and sharing of their knowledge about certain places (navigation points and orientation cues).

1.3 User study

Extensive research studies have already been carried out in the field of cognitive psychology, especially research on cognitive mapping, way-finding and the coding strategies of visually impaired persons [11,16,17,22]. This knowledge is essential for designing any navigation system for visually impaired persons that is based on navigation instructions and speech output.

Theoretical and empirical attempts to determine how the spatial abilities of blind persons compare with the abilities of the sighted population have led to the formulation of three main theoretical frameworks [22]. The empiricist 'deficiency theory' rejects any spatial understanding of congenitally blind persons due to the absence of visual experience. However, this theory has been found irrelevant due to lack of empirical evidence. The 'inefficiency theory' states that the spatial abilities of the blind are similar to those of sighted persons, but that there is worse performance. The 'difference theory' claims that the blind use qualitatively different cognitive strategies [16], and use different navigation points and orientation cues than sighted people. The difference theory leads us to the hypothetical implication that blind persons may be able to provide other blind persons with the right instructions, in line with the cognitive strategies that they use.

For example, the research carried out by Bradley and Dunlop [7] supports our idea with the finding that visually impaired people are more efficient and have a lower workload if they are given navigational instruction by other visually impaired people, rather than by sighted persons who are not trained in navigating the visually impaired. However, this work does not cover communication on navigation among visually impaired persons, their habits and their subjective preferences during the process of searching for help.

Further important research has been conducted in the field of the mental models that visually impaired persons form about the environment in which they are moving. It has been stated that different coding strategies for describing an environment are used by sighted persons and by visually impaired persons. The visually impaired tend to segment their description of the route, and make use of significantly more information about the route than sighted persons. Visually impaired persons also tend to use egocentric (body-centered) spatial references rather than external spatial references, which are more widely used by sighted persons [11,22]. The environment description provided by visually impaired persons is coded by the mental model that is also used by other visually impaired persons, and which is fundamentally different from the mental models formed by sighted persons. For this reason, the description provided by one visually impaired person for another should be better than the description provided by an untrained sighted person.

The principal problem is how to gather a specific description of the environment where visually impaired users will navigate. As a solution to the situation described in Fig. 2, we propose a navigation system based on collaboration among visually impaired users.

We have defined two essential conditions for operating a navigation system based on collaboration among visually impaired persons:

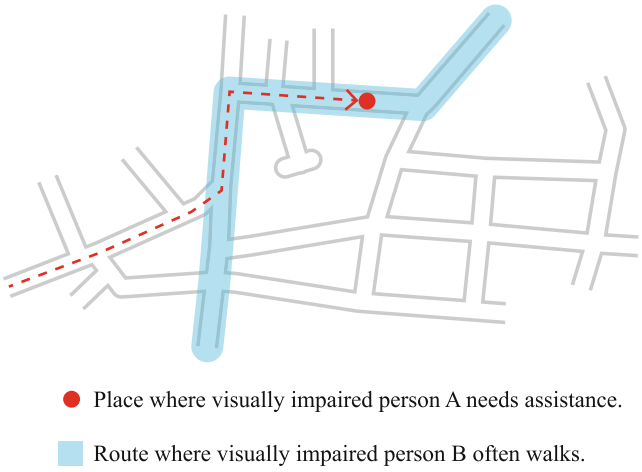


Fig. 2 Scheme illustrating the use case (see Sect. 1.1) and showing an environment description shared between visually impaired persons. The *blue area* marks the part of the urban environment known by visually impaired user B. Visually impaired user A has got lost in an area known to user B. The goal of the collaborative navigation system is to identify user B, and to connect user A to user B so that he/she can provide a description of the area and solve the navigation problem

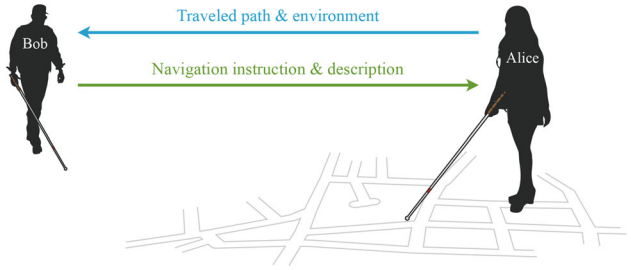


Fig. 3 Resuming from loss of orientation and getting navigation instructions: user A (Alice) shares her traveled path and a description of her surroundings with user B (Bob). Then she receives navigation instruction with a description of navigation points and orientation cues and further route instructions

- the existence of communication among visually impaired users about navigation, and the ability to share their knowledge (needed for helping other visually impaired users with navigation),
- and a sufficient length of regularly walked routes that they remember in every detail.

The main difference from existing state-of-the-art navigation systems (see Sect. 1), which are based on a description of the environment limited both in coverage of the city and by the number of navigation instructors, is the utilisation of knowledge of suitable quality distributed among a large number of visually impaired users (see Fig. 3) (common mental models, navigation points and orientation cues).

The proposed navigation system is based on two main principles:

- firstly, knowledge obtained from the cognitive psychology of visually impaired people (mental models of the environment), and
- secondly, knowledge provided by their natural behaviour (segmenting a route rich in navigation points and orientation cues recognised by visually impaired persons).

Collaboration among users can be generalised as cognitive information communication, in which communication takes place at cognitive level in intra-cognitive sensor sharing (i.e. members of the communication share what they perceive via several sensory channels. One visually impaired person describes what he/she hears, smells, touches, etc. to another) [4].

The problem of identifying and selecting a suitable visually impaired person who will provide a useful description of the environment for the lost user is not trivial, as has been shown above (see the Sect. 1.1). The complexity of this problem leads to the problem of reality (data) mining [10].

In order to explore the possibilities of setting up and utilising a collaborative navigation system of visually impaired persons, we conducted two studies.

The first study was conducted to clarify attitudes and habits of visually impaired users in communicating with navigation instructors, friends, family and strangers in the situation where they have lost their way or need help with navigation. The existence of communication among visually impaired people is the key condition for successful function of the navigation system.

The goal of the second study was to gain an insight into the structure and the length of the regular routes that visually impaired persons walk in the urban environment, and to estimate the average coverage of the city by their mental map and their ability to navigate others along the routes. It is obvious that there are individual differences in the distance and in the quality of the environment description (the level of detail of the environment that they walk in) that visually impaired people remember, or store in mental maps. This information is important for estimating the minimum number of users needed for successful operation of the proposed navigation system based on collaboration among visually impaired users.

The target group for all studies comprised visually impaired people of different ages, different duration, category and onset of disability and different technical skills. The target group consisted of persons with category 4 and category 5 blindness in the ICD-10 WHO classification [25]—blindness with light perception (category 4) and with no light perception—e.g. complete blindness (category 5).

2 Communication in the navigation of visually impaired

A user study was carried out in order to gain an initial insight into the question of communication in the navigation of visually impaired users. The study consisted of a qualitative part and a quantitative part [5].

- The qualitative study consisted of five semi-structured interviews, and was conducted to gain an insight into the problem and to form basic hypotheses for the subsequent quantitative study.
- The quantitative study was conducted via an e-mail based questionnaire. The aim was to obtain statistically valid data and to validate the hypotheses formulated during the qualitative study.

The recruitment of visually impaired people who will agree to participate personally in a study is rather problematic. For the quantitative study we therefore selected participants from a group of visually impaired persons who have already been collaborating with our department for a considerable period of time.

The quantitative study was conducted via email through the mailing list of the navigation and education centre

for visually impaired people of Czech Blind United—SONS [20].

2.1 Qualitative study

The qualitative study consisted of five 1-h semi-structured interviews with visually impaired participants. The goal of the qualitative study was to obtain in-depth information about participants favourite means of communication with other people, their openness to communicating with strangers, their willingness to help them with navigation, and the privacy problem of the location sharing and storing.

The topics for the interviews were: how often the participants seek help from strangers, their feelings in stress situations, their favourite narrative style for describing a route, and their willingness to help other unknown visually impaired people in navigating, and to participate in a programme for collaborative navigation. When evaluating the qualitative research results, it is necessary to consider the potential bias due to the sensitivity of questions dealing with participants feelings and stress.

The following topics were discussed with five participants between the ages of 25 and 64 years (4 males, 1 female) in the semi-structured interviews:

- the community of visually impaired persons,
- collaboration in navigation,
- experiences with location services,
- privacy issues of location sharing,
- behaviour in situations of loss of orientation.

Table 1 contains key factoids collected during the qualitative study with visually impaired participants. Most of the participants had experience of navigating other visually impaired persons, and/or had themselves been navigated by another visually impaired person (No. 1). Most of the participants assessed instructions given by a visually impaired person as better than the instructions given by sighted persons (No. 2). Most of the participants consider asking strangers for directions to be a natural part of the navigational process (No. 3). More than half of the participants expressed concerns about sharing their location for the purposes of helping other visually impaired persons (No. 4).

Table 1 Key factoids collected during the qualitative study

No.	Factoid	Occurrence
1	Participant has experience of navigating other visually impaired persons, and has been navigated by other visually impaired persons by mobile phone, ICQ, etc	5
2	Navigational instructions given by a visually impaired person about a place that he/she knows place are better than instructions from a sighted person	4
3	Asking strangers in the street for directions is a natural part of the navigation process	4
4	Participants have concerns about location sharing	3

One of the participants also mentioned that there is a problem with finding someone suitable to ask for help—some strangers do not want to help him/her, or do not speak Czech—and for this reason he/she does not like to ask people in the street.

2.1.1 Hypotheses

On the basis of the qualitative study, we formulated the following hypotheses:

- H1: A visually impaired user can navigate another visually impaired user via mobile phone.
- H2: A visually impaired user will prefer navigation instructions provided by another visually impaired person to instructions from a sighted person.
- H3: A visually impaired user will not hesitate to ask strangers in the street in order to get reliable directions.
- H4: A visually impaired user will allow information to be collected about his/her location in order to help in navigating other visually impaired people.

2.2 Quantitative study

The quantitative study was conducted on the basis of a questionnaire compiled from the findings obtained in the qualitative studies. The goal was to validate the hypotheses formulated on the basis of the qualitative study (see the section above). The questions were focused on the following topics:

- behaviour while obtaining navigation information from strangers in the street,
- preferences in selecting contacts when seeking for navigational instruction,
- navigating friends and family members via phone, email, instant messaging, etc.,
- requesting help with navigation from friends or family members,
- willingness to help an unknown visually impaired person with navigation,
- privacy problems of location sharing.

Fifty-four visually impaired respondents aged from 20 to 80 years took part in the quantitative study. The respon-

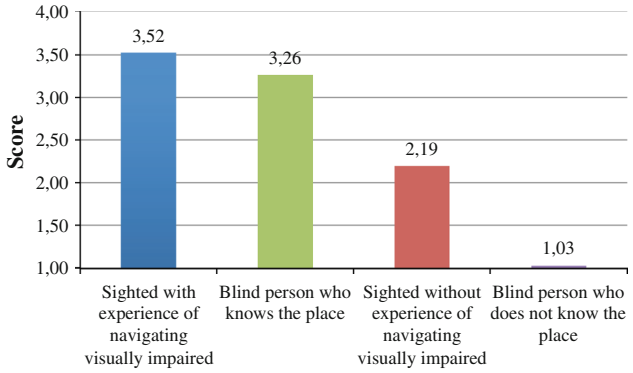


Fig. 4 Preferences in selecting a contact for help in navigating (n = 31)

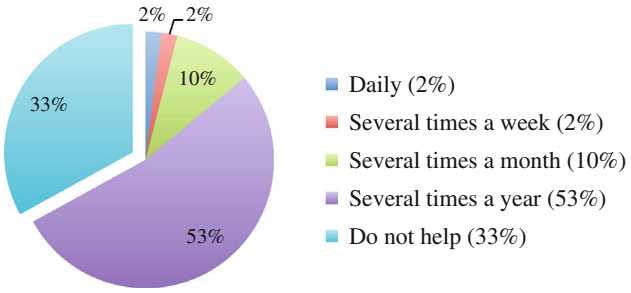


Fig. 5 Percentage of respondents helping friends or family members to navigate (n = 54)

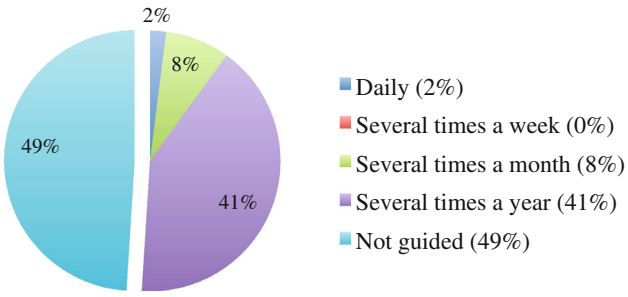


Fig. 6 Percentage of respondents who are guided by friends or by family members (n = 54)

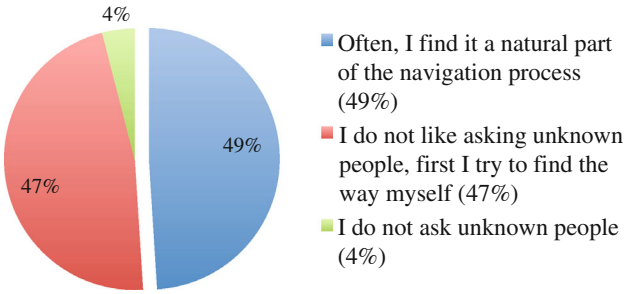


Fig. 7 Asking strangers in the street for directions (n = 54)

dents were 36 males and 18 females, with an average age of 47 years. 17 of the respondents were congenitally blind.

The questionnaires were sent by email to the mailing list of the navigation and education centre of Czech Blind United—SONS [20]. Not all of the responses were correctly filled in, due to the of type of impairment the respondent (if only some of the participants were able to fill the answer correctly, the exact number is mentioned in the results), and only the correctly filled in responses were taken into the account.

Figure 4 shows the preferences in selecting the contact who will help the visually impaired person to navigate in an unknown place. The higher the bar, the higher the score (minimum = 1, maximum = 4) calculated from the ordering of the following four variants: sighted person with experience of navigating visually impaired persons (score = 3.52); blind person who knows the place of navigation (score = 3.26); sighted person with no experience of navigating visually impaired persons (score = 2.19); and blind person who does not know the place of navigation (score = 1.03).

Most of the respondents selected a sighted person with experience in navigating visually impaired persons as the best option for getting good navigational information (navigation instructor). The important fact is that the respondents selected another visually impaired person as the second best option for getting good navigation information, rather than a sighted person with no experience of navigating visually impaired persons.

Due to the complicated instructions on how to answer the question from which Fig. 4 emerges the ordering of 4 variants—complete responses were collected from only 31 respondents.

Figures 5 and 6 show the behaviour in a situation when the visually impaired user helps or is guided by another visually impaired friend or family member. The majority i.e. 67 % of the respondents (daily 2 %, weekly 2 %, monthly 10 %, yearly 53 %) have experience of navigating friends or family members by phone, by email or by instant messaging, or was guided themselves i.e. 51 % of the respondents (daily 2 %, weekly 0 %, monthly 8 %, yearly 41 %).

About one half of the respondents (51 %) do not like asking strangers for help or do not ask them at all (47 %, do not ask; 4 % do not like to ask). However, the remaining 49 % consider asking strangers in the street to be a natural part of the navigation process (see Fig. 7).

Figure 8 shows that 68 % of all respondents have no concerns about sharing their location in order to help a visually impaired user in need of help to find the right contact. Only 12 % of the participants (mind 4 %; mind to some extent 8 %) would have problems with location sharing.

2.3 Discussion

It seems that hypothesis H1 is proven, as the visually impaired users reported that they can navigate or be navigated by other

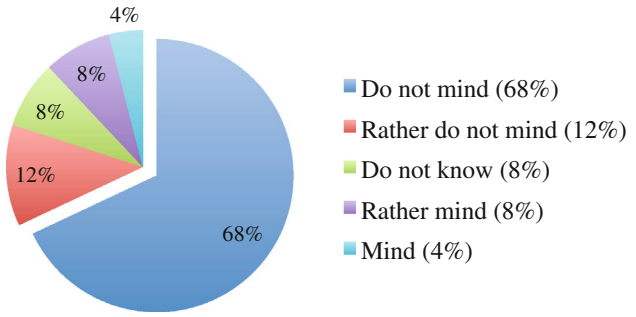


Fig. 8 Percentage of respondents concerned about location sharing (n = 54)

visually impaired users. Figures 5 and 6 show that visually impaired users already collaborate on navigating in small groups of users—family or groups of friends. We believe that appropriate tools and aids could further enhance the collaboration of these groups. These tools and aids should help to connect visually impaired users who do not know each other but can collaborate in navigating.

Hypothesis H2 is valid if we only take into consideration untrained sighted persons. H2 is not valid for expert sighted persons. Figure 4 shows that visually impaired people rate the quality of the description provided by a sighted expert more highly than the description provided by a visually impaired person who knows the area.

Respondents seem to expect that the description provided by a sighted expert will be more customised to their needs, as the expert has more information sources available.

The validity of hypothesis H3 is not proved. 51 % of the participants hesitate to ask strangers for instructions for navigation (see Fig. 7). However, the size of this group of participants can be attributed to a fact mentioned by one of the participants in the qualitative study—the problem of finding a suitable person who is willing to help. We believe that this problem can be solved by implementing our future system, where all of the users proposed by the system to provide help will have valuable information for navigation and will be willing to help.

In order to select accurately a suitable person to provide the right information for a visually impaired person who is lost, the location of all users needs to be stored and analysed. As is shown in Fig. 8, 68 % of all respondents have no problem with location sharing and location storing. If we add in those participants who do not mind only to some extent (12 %), a total of 80 % of the participants provide support for hypothesis H4.

3 Regular routes of visually impaired

We conducted a second study aimed at getting an insight into the routes that visually impaired people often walk in an urban environment, and at measuring the average route length per participant to estimate the minimum user base for our proposed navigation system based on collaboration among visually impaired persons.

A *regular route* is a route that a visually impaired person walks very often (e.g. weekly) and is able to describe in great detail with leading lines, navigation points and orientation cues recognisable by visually impaired persons. These routes can typically lead from home to the nearest public transport stop, shop, school, work, etc.

Visually impaired persons were recruited via the mailing list of the navigation and education centre for visually impaired people of Czech Blind United—SONS [20].

3.1 Method

We first considered long-term GPS tracking as a method for collecting regular routes of visually impaired people. As we are interested in regularly traveled routes, we planned to track each person for a period of approximately 4 weeks. Then we would make an interview to verify the quality of the environment description remembered by the visually impaired participants.

On the one hand, a long-term field study would provide more data for the measured regular routes, and would, for example, enable us to observe how long it might take to learn a new route. On the other hand, a long-term field study would require additional interviews to find out the level of detail of the stored description of the environment (and also to prove that the new routes had been learned), and it would require a long time for observations. This would prolong the experiment excessively, in view of the number of participants needed. There would also be a problem with removing data gathered in public transport from the regular routes, as we are only interested in the routes that visually impaired people walk and the mental model of the environment that they create. There would also be problems with coping with the poor accuracy of GPS in an urban environment (possible errors in tens of meters). In addition, the participants would have to be trained in operating tracking devices, e.g. charging batteries and carrying out basic maintenance tasks.

Considering all pros and cons of the method discussed above, we decided to choose a qualitative approach. This would enable us to explore regular routes in greater detail (e.g. which side of the street the participant walks on). We invited 20 visually impaired participants with category 4 blindness and category 5 blindness on the ICD-10 WHO classification scale [25] (category 4: blindness with light perception, and category 5: blindness with no light perception) to discuss regularly traveled routes in an urban environment.

There were ten males and ten females, average age 42.3, average duration of disability 26.3 years. Eight out of the 20 participants had category 4 blindness, while 12 had category

Table 2 List of participants in the study, with their level of disability and their age

No. of part	Male/female	Cat. 4/5	Mean age	Mean dur. of dis.	Cong./late blind
20	10/10	8/12	42.3	26.3	9/11

Table 3 Results of the study of regularly walked routes

No. of part	Mean length of the routes	Total length of the routes	Mean number of dest.
20	4.6 km/part	93.3 km	4.2/part

5 blindness [25]. Nine participants were congenitally blind, and 11 participants were late blind. Two participants used a guide dog. Sixteen participants in the study lived in Prague, and the four remaining participants lived in other large towns (see Table 2).

The main objectives of the interviews were to:

- identify regular routes walked by the visually impaired participants,
- measure the approximate length of the regular routes,
- estimate the contribution of the regular routes walked by the visually impaired persons to the current network of public transport,
- estimate number of participants needed to ensure successful operation of a collaborative navigation system.

The participants were asked to describe their weekly walked routes, e.g. to their work, shops, library, the nearest public transport stop, etc., in as great detail as possible, as if they were navigating another visually impaired person. Later we reconstructed the routes from the recorded description on to a map, measured the length of the walks for each participant, and created a heat map (see Figs. 12, 13). We included both directions for some routes (e.g. from home to the bus stop and back), as the description of the route can differ significantly in the leading lines, the slope of the surface, and other navigation points and orientation cues.

3.2 Qualitative study

The 20 visually impaired participants who were interviewed walk a total of 93.3 km of regular routes, and are able to describe the environment of these walks (see Table 3). This sum represents the length of unique regular routes for all participants (e.g. the route to the shop twice a week counts as one unique regular route). The mean length of the routes per participant is 4.6 km. The participants in the study visit 4.2 destinations, on an average (e.g. restaurants, workplaces, schools, hobbies, shops, etc.). The nearest public transport stops from home are not counted as destinations if they lie on the way to other places.

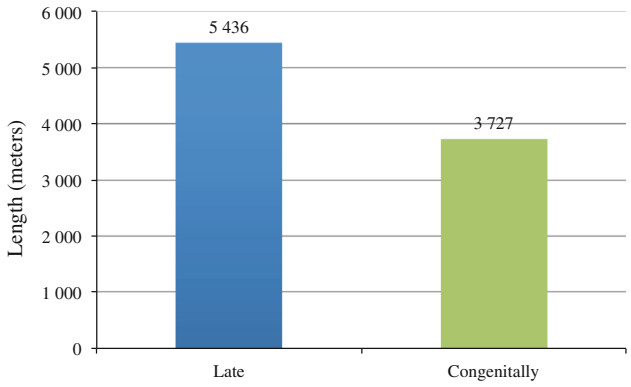


Fig. 9 Effect of the onset of blindness (congenital blindness or late blindness) on the mean length of the regularly walked routers (n = 20)

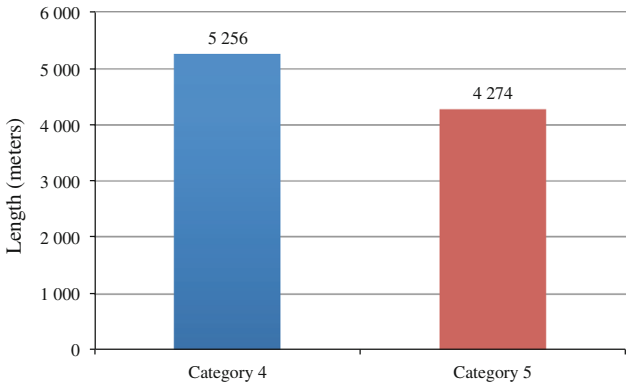


Fig. 10 Effect of the category [25] of blindness on the mean length of the regularly walked routes (n = 20)

Figure 9 shows the effect of congenital blindness and late blindness on the length of the regularly walked routes. There is a significant difference in the mean length of the routes for the late blind group, whose members walk 1,709 m further than the congenitally blind group. We think this may be partly due to inertia from the time when the participants were sighted, and when they had greater mobility and traveled more around the city. On the other hand, the group of congenitally blind persons tends to optimise from birth, and to use the best (safest) route. They also tended to use public transport as much as possible. However, this assumption needs to be further investigated in a study of the different cognitive mapping strategies employed by congenitally blind and late blind persons.

Figure 10 shows the relation between different categories of visually impairment and the length of the regularly walked routes. The participants with blindness with light perception (category 4) had on an average 982 m longer regular routes than participants without light perception. This may be because visually impaired participants with category 4

Table 4 The effect of the combinations of onset and category of blindness on the mean length of the regularly walked routes

	Category 4	Category 5
Congenitaly blind	5.3 km/part	2.5 km/part
Late blind	5.2 km/part	5.6 km/part

blindness use the remains of their sight (light perception) as a navigation aid, or because they have a deteriorating impairment and still retain many visually-oriented memories of the environment. Table 4 shows the mean length of the regularly walked routes for combinations of different onsets and categories of blindness.

Figure 11 shows data for all participants, together with their age. As expected, there are significant differences among the participants. Some of the participants walk long routes to the nearest public transport, or just for recreational walking. On the other hand, other participants avoid walking and use public transport as much as possible, or use a guide for unknown or complicated places.

Figure 12 shows a heat map, which maps a number of routes in a certain place from blue to red values. Red areas indicate places where there are larger numbers of regularly walked routes involving one participant or more. Similarly, Fig. 13 shows the same heat map for the whole city of Prague.

3.3 Discussion

Many of the regular routes of the participants start or end at public transport stops, which are usually easily accessible for visually impaired people. The total length of the public transport network in Prague is 1,029.8 km (subway 59.4 km, tram 142.4 km, bus 828 km) [9]. The regular walking routes taken by the 20 visually impaired persons who participated in the study have a combined total length of 93.3 km. These walking routes extend the current public transport network by 9 %, and create access to areas around public transport stops.

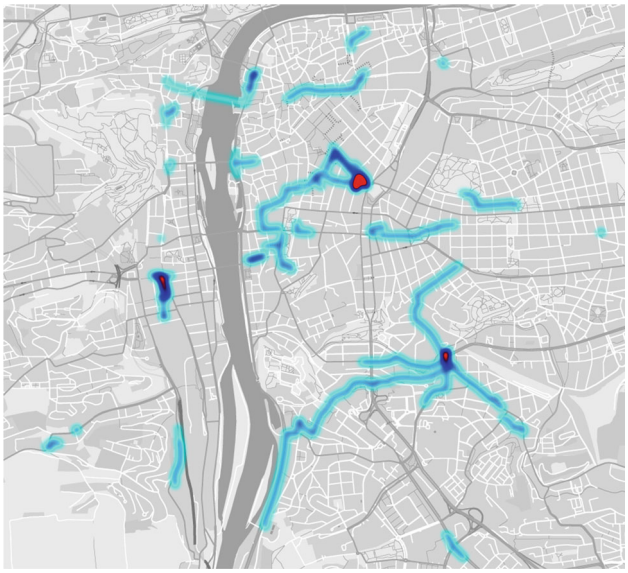


Fig. 12 Heat map of routes in city center of Prague

To cover the area accessible nowadays by public transport, the proposed collaborative navigation system for visually impaired persons would require approximately 200 active users for Prague (with none of them sharing the same route). The Blind United Union (SONS [20]) has over 10,000 members in the whole Czech Republic. Considering that one tenth of the population lives in Prague, there will be at least 1,000 visually impaired persons in the city. This number corresponds with the estimated number of 10,000–20,000 visually impaired (blind) persons in the Czech Republic. These numbers meet the essential condition there should be a sufficient length of regular routes to make the collaborative navigation system feasible.

4 Conclusion

The results from the first study have revealed that most of the hypotheses about communication among visually impaired

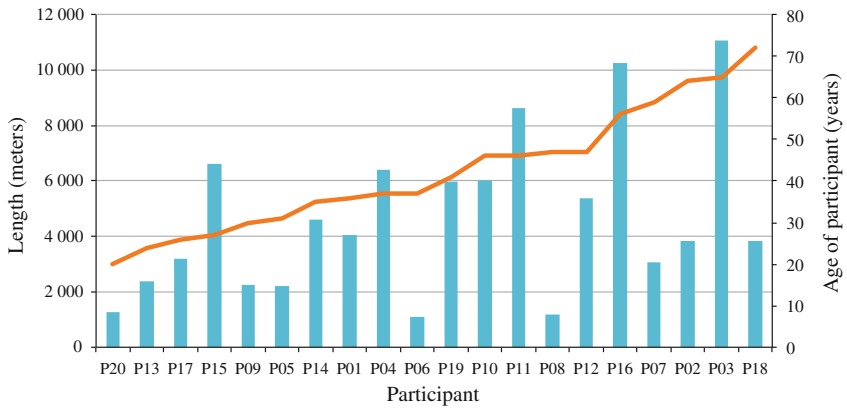


Fig. 11 Relation between the age of the participant (*orange line*) and the length of the regularly walked routes (*blue column*) (n = 20)

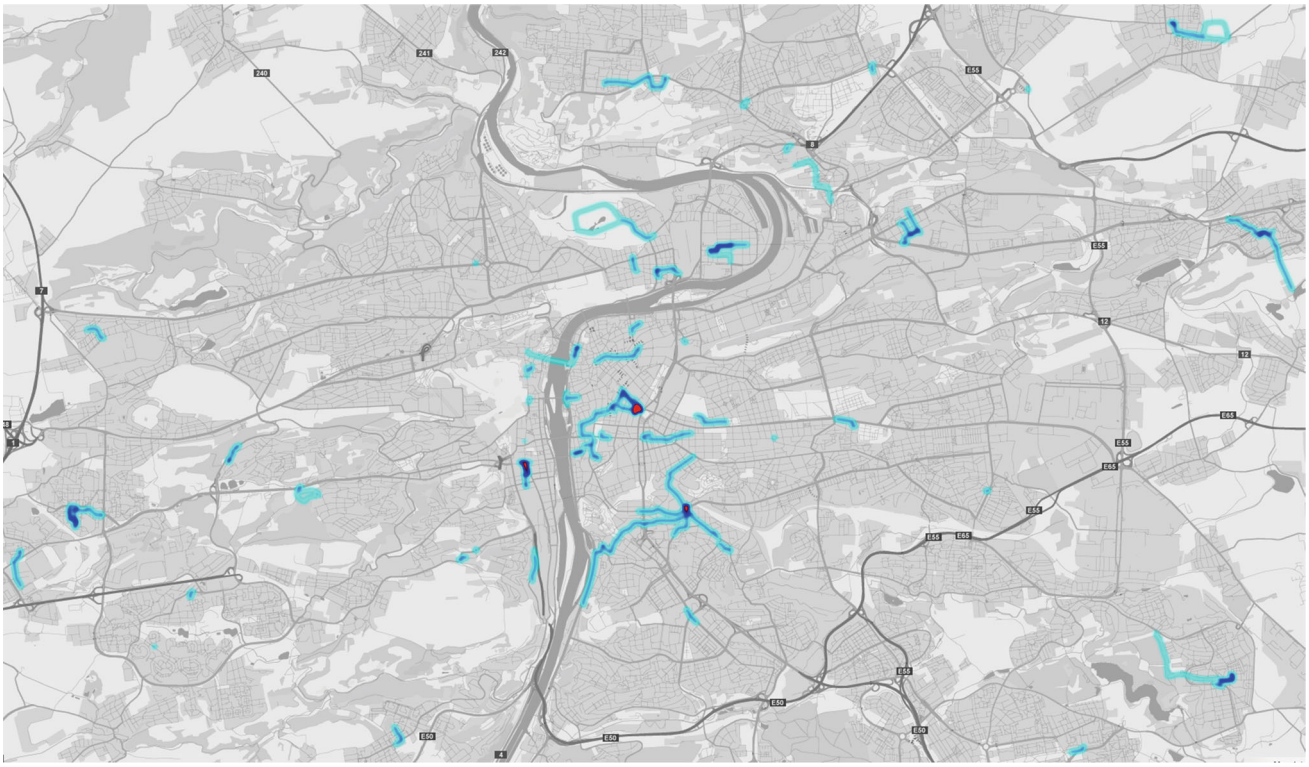


Fig. 13 Heat map of routes in whole city of Prague

people during navigation were correct, and provided important insights into the problems (see Sect. 2).

The data collected from the second study on regularly walked routes showed that visually impaired persons remember quite long route descriptions (in cognitive maps) and that they are able to describe them to other people with all important navigation points and orientation cues (see Sect. 3).

In summary, our studies have shown that both communication in navigation and regular routes of visually impaired people exist in suitable amounts and in suitable quality. These were the essential conditions for attempting to set up a navigation system based on collaboration among visually impaired persons. Both essential conditions have been fulfilled, and it has been shown that successful operation of the system is feasible.

The navigation system based on collaboration requires the collection of data on regular routes for future analysis, and the selection of visually impaired people who can provide help with navigation. Pilot testing and data collection should therefore be set up. The selection of participants suitable for the pilot testing will be based on their technological skills, as they will be required to interact with an early prototype of the data collecting application. The algorithm for recommending appropriate contacts for help will be created on the basis of the data collected from the pilot testing.

We are also currently focusing on the behaviour of visually impaired people in situations when they contact another unknown visually impaired person for help in navigation. A field study is currently running with participants who are intentionally navigated with wrong instructions and therefore get lost. The goal of this study is to observe the navigation points and orientation cues that are used by visually impaired persons who are lost to describe their situation, and the typical course of a conversation with another unknown visually impaired person.

Acknowledgments This research has been supported by the project Design of special user interfaces funded by Grant No. SGS13/213/OHK3 /3T/13 (FIS 161—832130C000) and it has been supported by the Technology Agency of the Czech Republic under the research program TE01020415 (V3C—Visual Computing Competence Center).

References

1. Android (2013) Introducing Android 4.0. http://www.android.com/about/ice-cream-sandwich
2. Apple (2013) Accessibility—VoiceOver. http://www.apple.com/accessibility/voiceover
3. Balata J (2011) System supporting tourism of visually impaired people. Master thesis, Czech Technical University in Prague, Faculty of Electrical Engineering, Czech Republic
4. Baranyi P, Csapó Á (2012) Definition and synergies of cognitive infocommunications. Acta Polytechnica Hungarica 9(1):67–83

5. Barkhuus L, Rode J (2007) From mice to men—24 years of evaluation in chi. In: Proceedings of the SIGCHI conference on human factors in computing systems. CHI07-Alt. CHI. ACM, New York
6. BlindSquare (2013) What is BlindSquare. http://blindsquare.com/about
7. Bradley NA, Dunlop MD (2005) An experimental investigation into wayfinding directions for visually impaired people. Pers Ubiquitous Comput 9(6):395–403
8. Code Factory (2013) Mobile speak. http://www.codefactory.es/en/products.asp?id=316
9. DPP (2013) Profil spolecnosti. http://www.dpp.cz/profil-spolecnosti
10. Eagle N, Pentland A (2006) Reality mining: sensing complex social systems. Pers Ubiquitous Comput 10(4):255–268
11. Golledge RG, Klatzky RL, Loomis JM (1996) Cognitive mapping and wayfinding by adults without vision. Geoj Libr 32:215–246
12. GPS (2013) Official U.S. Government information about the global positioning system (GPS) and related topics. http://www.gps.gov
13. Hunaiti Z, Garaj V, Balachandran W, Cecelja F (2005) Use of remote vision in navigation of visually impaired pedestrians. In: Jones S (ed) International congress series, vol 1282. Elsevier, London, pp 1026–1030
14. Loadstone GPS (2013) Navigation for blind mobile phone users. http://www.loadstone-gps.com
15. Microsoft (2013) Kinect for Windows. http://www.microsoft.com/en-us/kinectforwindows
16. Millar S (1994) Understanding and representing space: theory and evidence from studies with blind and sighted children, vol 198521421. Clarendon Press, Oxford
17. Millar S (1995) Understanding and representing spatial information. Br J Vis Impair 13(1):8–11
18. Prochazka O (2011) System for creating text description of route for visually impaired people. Master thesis, Czech Technical University in Prague, Faculty of Electrical Engineering, Czech Republic
19. Ran L, Helal S, Moore S (2004) Drishti: an integrated indoor/outdoor blind navigation system and service. In: Proceedings of the second IEEE annual conference on pervasive computing and communications (PerCom 2004). IEEE, pp 23–30
20. SONS (2013) Czech Blind United. http://www.sons.cz
21. Symbian Foundation (2013) Symbian. http://licensing.symbian.org
22. Ungar S (2000) Cognitive mapping without visual experience. In: Kitchin R, Freundschuh S (eds) Cognitive mapping: past, present, and future, vol 4. Routledge, London, pp 221–248
23. Völkel T, Kühn R, Weber G (2008) Mobility impaired pedestrians are not cars: requirements for the annotation of geographical data. In: Miesenberger K, klaus J, Zagler W, Karshmer A (eds) Computers helping people with special needs. Springer, New York, pp 1085–1092
24. Vystrcil J (2008) User interface for in-door navigation of visually impaired people via mobile phone. Master thesis, Czech Technical University in Prague, Faculty of Electrical Engineering, Czech Republic
25. WHO (2013) International classification of diseases v. 10. http://apps.who.int/classifications/icd10
26. Zöllner M, Huber S, Jetter HC, Reiterer H (2011) NAVI—a proof-of-concept of a mobile navigational aid for visually impaired based on the Microsoft Kinect. Springer, New York

Appendix D

Hands Free Mouse: Comparative Study on Mouse Clicks Controlled by Humming

Polacek O., Mikovec Z.: Hands Free Mouse: Comparative Study on Mouse Clicks Controlled by Humming. In Proc. of the 28th of the int. conference extended abstracts on Human factors in computing systems. New York: ACM, 2010, p. 3769-3774. ISBN 978-1-60558-930-5.

Hands Free Mouse: Comparative Study on Mouse Clicks Controlled by Humming

Ondřej Poláček
Faculty of Electrical Engineering,
Czech Technical University in
Prague
Karlovo nám. 13
12135 Praha 2
Czech Republic
polacond@fel.cvut.cz

Zdeněk Míkovec
Faculty of Electrical Engineering,
Czech Technical University in
Prague
Karlovo nám. 13
12135 Praha 2
Czech Republic
xmikovec@fel.cvut.cz

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-930-5/10/04.

Abstract
In this paper we present a novel method of simulating mouse clicks while the cursor is navigated by head movements tracked by webcam. Our method is based on simple hummed voice commands. It is fast, language independent and provides full control of common mouse buttons. Our method was compared with other three different methods in an experiment that proved its efficiency by means of task duration.

Keyword
Non-Verbal Vocal Interface, Voice Interface, Head Tracking, Accessibility, Comparative Study

ACM Classification Keywords
H.5.2 Information interfaces and presentation: User Interfaces – Input devices and strategies; Voice I/O.

General Terms
Design, Experimentation, Measurement, Performance

Introduction
Complex graphical user interfaces (GUI) are present not only in desktop computers, but they also appear in other areas such as Rich Internet Applications (RIA) on the Internet. The efficiency of interaction with such complex GUI is strongly dependent on the efficiency of

the way the mouse is used. For users with limited motor abilities (especially upper-limb impaired users) the use of mouse could be a serious problem. These users need an alternative way of mouse control. There are several hands-free mouse solutions which can solve the problem of cursor control satisfactorily, but not the problem of simulation of mouse clicks (see State of the Art). The set of the mouse clicks simulated is either not complete or too complicated for the user who cannot simulate mouse in full extent.

State of the Art

Several methods have been used to simulate mouse clicks. The camera mouse system [1] used *a dwell time* method. A mouse click was generated, when the user kept the mouse cursor within a 30-pixel radius for 0.5 s. The dwell time method is capable of simulating left click only and does not cope with other mouse events such as right click, double click, dragging and scrolling. Moreover the method raises the *Midas touch problem* [4], as the user cannot stop the cursor without issuing a left click. It can be solved either by adding places where the user can stop the cursor [1] or displaying a pop-up menu after the dwell time expires [11].

Tracking various face features can be also used to simulate mouse clicks. For instance, a system published by Tu [10] responded to the state of user's mouth. It was capable of simulating left click by opening the mouth and right click by stretching the mouth. Dragging was provided by moving the cursor while keeping the mouth open. Another system called hMouse [2] triggered left and right clicks when the turn of user's head exceeded a specific threshold angle. There was no solution for dragging and scrolling reported.

There are also several multimodal systems that use different interaction channel to simulate mouse clicks. Nouse [3] employed a computer keyboard, which is a rapid and complex solution, however, the users still need to use their hands and such solution cannot be used for disabled people with severe upper-limb impairment. Another modality that can be used for simulating mouse clicks is speech. Multimodal system published by Loewenich and Maire [5] defined five simple speech commands (*click*, *double*, *right*, *hold* and *drop*) that covered all clicking and dragging operations. Ronzhin and Karpov [8] published similar system that defined 30 speech commands covering clicking, dragging and scrolling one by one line. Remaining commands were used as shortcuts to common operations such as *open a file*, *exit an application*, *copy*, *paste* etc.

Our Solution

In our system *non-verbal vocal interaction* (NVVI) [6, 9] is used for simulating mouse clicks. This interaction method can be characterized as using other sounds than speech, such as humming, to control user interfaces. In our case, hummed voice commands are determined by its pitch and length. Expected pitch profiles of the commands are depicted in Figure 1. Left click (1a) is defined as a short tone produced below user-specific threshold pitch. Double click (1b) is defined as two consecutive left clicks. Right click (1c) is a short tone above the threshold pitch. Drag (1d) is a long tone. The difference between long and short tone is 0.5 s. However, this value can be modified according to preferences of the user. Drop operation does not

have its own command and it is triggered by short or long tone. Scrolling (1e, 1f) is performed when significant increase or decrease in pitch is detected. Amount of lines scrolled is determined by length of the voice command in real time. Continuous real-time control is a significant advantage of NVVI [9]. Using speech the user has to explicitly specify the amount of lines scrolled, which is rather awkward. Note that in order to keep our method simple, minimal amount of voice commands was used. Moreover, commands are very simple and short for the most frequent operations (clicking) and they are a bit more complicated for advanced operations (dragging and scrolling).

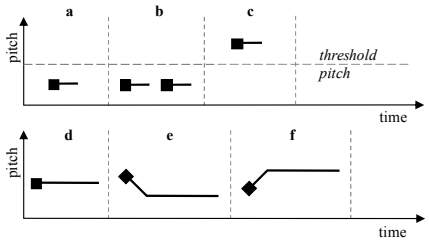


Figure 1. Non-verbal vocal commands used to simulate mouse clicks. **a**. left click, **b.** double click, **c.** right click, **d.** drag, **e.** scroll down, **f.** scroll up

This method is very well suited for real-time control, as the hummed commands are recognized much faster than verbal commands [9]. They are also culturally and language independent. On the other hand this is a very unusual way of interaction and the users have to get used to it [6].

Experiment

The aim of the experiment was to determine the efficiency of our solution. We compared our NVVI method with other three different methods for simulating mouse clicks in terms of speed and error rate. The following four methods were prepared for the comparison test:

- *Non-verbal vocal interaction* (NVVI) as described in previous section. Scrolling voice commands were not included in the experiment.
- *Speech commands*. Regarding the fact that all participants were Czech native speakers we used Czech commands recognized by MyVoice application [7]. Commands and their English equivalents are listed in Table 1.

Table 1. Speech commands.

Command in Czech	English Equivalent	Description (mouse operation)
Klik	Click	Left click
Dvojklik	Double Click	Double left click
Pravý klik	Right click	Right click
Vzít	Drag	Left button down
Položit	Drop	Left button up

- *Computer keyboard*. Mouse buttons were mapped to keystrokes. Alt + left arrow corresponded to left mouse button and alt + right arrow to right button. Arrows up and down corresponded to mouse wheel. This method was chosen as a reference test.
- *Head gestures*. This solution combined the dwell time approach with a pie menu (see Figure 2). When the mouse cursor did not significantly move for 0.5 s, the pie menu appeared and concrete operation was chosen by moving the cursor over the menu as

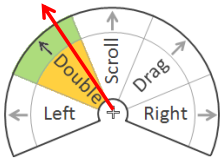


Figure 2. Pie menu.

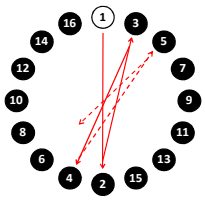


Figure 4. Task template.

depicted in Figure 2 by red arrow. This method does not suffer from Midas touch problem [4] and all mouse operations can be simulated. The menu can be cancelled by moving the cursor down.

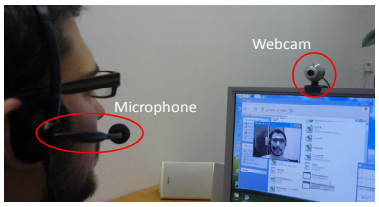


Figure 3. Setup of the experiment.

In order to navigate the cursor by head throughout the experiment, a head tracking system described in [11] was chosen. A cheap webcam can be used by the system to track the head of the user and convert its position and turn to position of the mouse cursor on a screen. An absolute mode is used as described in [5], i.e. the position of head is directly mapped to the position of the mouse cursor. A 17" LCD monitor with native resolution 1280 x 1024 pixels was used. There was a webcam mounted on the top of the monitor that provided data for head tracking. While using vocal modalities the participants used headphones with microphone. The experiment setup is depicted in Figure 3.

For the experiment four mouse click simulation tasks were defined. In every task the participants had to move the cursor to particular circle as shown in Figure 4 and perform specified clicking operation. The participants had to start with circle 1 and continue until circle 16 was reached. Part of expected cursor trajectory is shown by red arrows. Every task defined a different mouse operation to be performed in the circles as follows:

- Task *Pointing*. No clicking was involved. This task was included for reference purposes.
- Task *Left Click*. Only left click had to be simulated.
- Task *Multi Click*. Left, right and double clicks were simulated according to caption of circles.
- Task *Drag & Drop*. Drag and drop operations were involved.

The participants had to pass through overall 16 tasks (four tasks using four modalities for simulating mouse clicks). In order to minimize learning effect, the sequence of methods and tasks was shuffled. Moreover every task had to be undertaken twice and data were measured only in the second try. The objective data collected were processed into three indicators as follows:

- *Task duration*, which is the duration between the first and last operation in a task including error operations. This indicator is used to measure the efficiency of each method.
- *Click duration*, which is duration between passing the border of a small circle and a correct click.
- Error rate, which expresses the number of wrong clicks relative to number of total clicks.

Due to the long-lasting single session (about 50 minutes), we did not include scrolling capabilities of evaluated methods. After each session the participants

Table 2. Mean times and standard deviations (SD) for each task and modality. Grey cells in one row correspond to means that are not statistically different. Overall error rates for each method are shown in the last row.

		Speech		**NVVI**		**Keyboard**		**Head gestures**	
		Mean Time [s]	SD [s]	Mean Time [s]	SD [s]	Mean Time [s]	SD [s]	Mean Time [s]	SD [s]
Left Click	Task Duration	49.6	7.4	39.0	7.8	34.0	6.5	60.9	9.7
	Click Duration	1.342	0.277	0.875	0.203	0.524	0.175	1.927	0.260
Multi Click	Task Duration	55.1	7.7	49.5	12.9	39.3	7.0	71.1	14.3
	Click Duration	1.639	0.374	1.358	0.526	0.706	0.213	1.977	0.415
Drag & Drop	Task Duration	51.0	6.7	44.5	8.2	36.3	6.8	70.5	9.9
	Click Duration	1.430	0.274	1.256	0.249	0.558	0.186	1.667	0.314
Error rate [%]		3.53		6.11		4.09		1.75	

were given a post-test questionnaire to subjectively assess speed, comfort and accuracy of each method. In the experiment 54 participants without disabilities took part. They were recruited from university students (mean age=23.5, SD=0.98) and were technically oriented and experienced computer users. The participants were trained to perform NVVI and head gestures in a training session which was conducted before the experiment and lasted approximately 15 minutes. Speech and keyboard methods were not trained because speech is a natural form of interaction and participants were experienced enough in using keyboard.

Results

In Table 2 the results of the experiment are summarized. Mean times (task and click durations) of each method in three tasks are shown in rows. As these times are compared in each row, the speed of the methods can be evaluated without exception as follows:

Keyboard < NVVI < Speech < Head gestures

ANOVA test and Scheffé's method were used to find statistically significant ($p < .01$) differences in mean times of each task. Most of them are significant except those shown in grey color in Table 2. Our method is the fastest among hands-free methods (speech and head gestures), however, it is slower than keyboard, which on the other hand is unusable for severe motor impaired users.

Error rate results are summarized in the last row of Table 2. The head gestures method experienced the lowest error rate. This is probably caused by the relatively high time penalty, when the user selects a wrong option. In this case the cursor has to be navigated to the initial position and the user has to wait for pie menu popup (dwell time). This leads to much more careful interaction. However, we believe, that this behavior can be improved by personalizing the dwell time and size of the pie menu. The error rate of NVVI was the highest one (6.11%), which is caused by insufficient training involving only one session. According to longitudinal studies [6, 9] four training sessions are enough to minimize error rate of the NVVI.

Nevertheless the time penalty caused by these errors is already included in the task duration indicator.

Table 3. Questionnaire results. Scale 1 (=worst) … 5 (=best). Mean values are displayed.

	Speech	NVVI	Kbd.	Gest.
Speed	3.33	3.50	4.72	2.11
Comfort	3.74	2.81	4.30	2.89
Accuracy	3.94	2.94	4.81	3.04

Subjective results are shown in Table 3. Head gestures were subjectively rated by the users as the worst method and keyboard as the best. Even though NVVI was faster than speech, it was perceived worse in comfort and accuracy.

Conclusion

In this paper we have described a method for mouse clicks simulation based on humming (NVVI). This method is capable of simulating all common mouse buttons including mouse wheel for real-time scrolling. Our method was compared with other three methods (speech, head gestures and keyboard) and it was the second fastest, although it experienced the highest error rate. The subjective perception of the accuracy and comfort was also rated as the lowest. In the future, we will conduct longitudinal tests with disabled users in real applications and combine more modalities in the system to provide more efficient control of a computer.

Acknowledgement

We would like to thank Lukáš Zich from the Center for Machine Perception, CTU in Prague for provision of the head tracking software. This research has been partially supported by the MSMT research program MSM 6840770014 and the VitalMind project (IST-215387).

References

[1] Betke, M., Gips, J. and Fleming, P. The Camera Mouse: Visual Tracking of Body Features to Provide Computer Access for People With Severe Disabilities. In IEEE Transactions on Neural Systems and Rehabilitation Engineering, IEEE Computer Society (2002), 1-10.

[2] Fu, Y., Huang, T. S., hMouse: Head Tracking Driven Virtual Computer Mouse, In Proc WACV'07, IEEE Computer Society (2007), 30-36.

[3] Gorodnichy, D. O., Malik, S. and Roth, G. Nouse 'Use Your Nose as a Mouse' - a New Technology for hands-free Gamers and Interfaces, In Proc VI'2002, Calgary (2002), 354-361.

[4] Jacob, R. J.K. What you look at is what you get, In Computer, vol.26, no.7, IEEE Computer Society (1993), 65-66.

[5] Loewenich, F. and Maire, F. Hands-free mouse-pointer manipulation using motion-tracking and speech recognition. In Proc OZCHI, ACM Press (2007),295-302.

[6] Mahmud, M., Sporka, A. J., Kurniawan, S. H. and Slavik, P. A Comparative Longitudinal Study of Non-verbal Mouse Pointer, In Proc INTERACT 2007, Springer-Verlag (2007), 489-502.

[7] Nouza, J., Nouza, T. and Červa, P. A Multi-Functional Voice-Control Aid for Disabled Persons. In Proc SPECOM 2005, Moscow, 715-718.

[8] Ronzhin, A. and Karpov, A. Assistive multimodal system based on speech recognition and head tracking, In Proc of EUSIPCO'2005, Turkey, 2005.

[9] Sporka, A. J., Kurniawan, S. H., Mahmud, M. and Slavik, P. Longitudinal study of continuous non-speech operated mouse pointer, In Proc CHI'07, ACM Press (2007), 2669-2674.

[10] Tu, J., Huang, T. and Tao, H. Face as Mouse Through Visual Face Tracking, In Proc CRV (2005), IEEE Computer Society (2005), 339-346.

[11] Zich, L. Video based Human-Computer interface, Master Thesis, 2009, CTU Prague, FEE.

Appendix E

Avatar and Dialog Turn–Yielding Phenomena

Kunc L., Mikovec Z., Slavik P.: Avatar and Dialog Turn–Yielding Phenomena. In Int. Journal of Technology and Human Interaction (IJTHI). 2013, vol. 9, no. 2, p. 66-88. ISSN 1548-3908.

Avatar and Dialog Turn-Yielding Phenomena

Ladislav Kunc, Department of Computer Graphics and Interaction, Czech Technical University in Prague, Prague, Czech Republic

Zdeněk Míkovec, Department of Computer Graphics and Interaction, Czech Technical University in Prague, Prague, Czech Republic

Pavel Slavík, Department of Computer Graphics and Interaction, Czech Technical University in Prague, Prague, Czech Republic

ABSTRACT

Turn-taking and turn-yielding phenomena in dialogs receive increasing attention nowadays. A growing number of spoken dialog systems inspire application designers to humanize people's interaction experience with computers. The knowledge of psychology in discourse structure could be helpful in this effort. In this paper the authors explore effectiveness of selected visual and vocal turn-yielding cues in dialog systems using synthesized speech and an avatar. The aim of this work is to detect the role of visual and vocal cues on dialog turn-change judgment using a conversational agent. The authors compare and study the cues in two experiments. Findings of those experiments suggest that the selected visual turn-yielding cues are more effective than the vocal cues in increasing correct judgment of dialog turn-change. Vocal cues in the experiment show quite poor results and the conclusion discusses possible explanations of that.

Keywords: *Dialog Systems, Embodied Conversational Agent, Human Computer Interaction, Turn-Taking, Turn-Yielding*

INTRODUCTION

Implementation of systems able to interact with humans in a natural conversational way to provide services, that would otherwise require communication by means of human phone operators or menu graphical-based systems, represents the ultimate goal of human-computer interaction designers. A voice based user interface is one of the possible human-computer interaction methods. Last few years have brought many advances in automatic speech recognition systems, text-to-speech systems, and in dialog management systems. These systems are getting more and more sophisticated and as such they have grown relatively complex. A good example of this trend could be automatic telephone systems which help users solve simple problems in daily life and, in difficult cases, reroute the customer to appropriate human operator. AT&T 'How may I help you?' system is credited to be the first among such advanced systems (Gorin, Riccardi, & Wright, 1997). However, computer

DOI: 10.4018/jthi.2013040105

spoken dialog technology is still far from being a near-human performance (Pieraccini, Suendermann, Dayanidhi, & Liscombe, 2009).

A speech based application should present a clear benefit over other styles of interaction. Spoken dialog applications are best when they enable something that cannot otherwise be done (e.g. safely operate a phone or a navigation system while driving); or when the user's hands and eyes are busy; or when the keyboard is not available. The usage of speech recognition systems could overcome some usability issues for specific groups of people: for example, tremor and age-related changes in bodily motor control represent a problem for navigating touch user interfaces; or people with different visual disabilities can hardly use visually based interfaces. Such groups of people could clearly benefit from speech recognition systems. Non-disabled users often prefer a menu and a "common" based style of interaction to spoken dialog systems (the common style of interaction means keyboard, mouse, touch based systems), e.g. in situations when the user is not in a private environment (Lai & Yankelovich, 2008). Another issue that people have with the spoken dialog systems is the way the system takes/yields a dialog turn. A human sends plenty of cues during a natural conversation to indicate a wish for turn-keeping or turn-yielding. However, spoken dialog systems use predominantly only one way to yield the turn nowadays, and that is the use of a long pause (Gravano, 2009), typically in the range of 0.5s to 1s (Ferrer, Shriberg, & Stolcke, 2002). But long pauses are not natural in human-to-human dialog; conversations in general tend to be smoother without them.

It should be taken into account that verbal (speech) communication is not the only part of interpersonal communication. The other part is nonverbal language (facial expressions, body gestures, etc.). While listening to others, people do not focus only on the verbal content of a conveyed message. During complex assessment of a speaking person we process both parts of speech: the nonverbal and the verbal (Hargie & Dickson, 2004).

Incorporating some form of face-to-face communication into spoken dialog technology could enable the system to express nonverbal parts of communication. Seeing virtual faces (that means conversational agents, avatars, etc.) also humanizes computer user interfaces and makes them more acceptable for common users (Yee, Bailenson, & Rickertsen, 2007). These so called embodied conversational agents (ECAs) or avatars, integrate gestures, facial expressions and visual / nonverbal aspects of speech into human-computer interaction (Cassel, 2000).

In this paper the issue of turn-yielding in two-party dialogs is addressed. The main motivation of the study is to make an ECA-based dialog application more natural and improve its interactivity by using turn-yielding mechanism. Various turn-yielding cues are applied and their impact on turn-change judgment is tested and compared.

ISSUES IN ARTIFICIAL DIALOGS

An overview of the current state of play in the field of turn-taking/yielding in spoken dialog systems is provided and research hypotheses are introduced here.

The primary goal of spoken dialog systems research is to produce a system that is capable of smooth conversation with a human in a specific domain. Implementation of such a system requires reliable speech recognition, dialog manager and speech synthesis. Although the recent quality improvement in such technologies is tremendous, dialogs with artificial systems still fall far behind in comparison with their human counterparts in terms of both comfort and efficiency. The reaction times of artificial dialogs are still slower, although the systems

are improved through incorporation of turn-taking models. The model used by Raux and Eskenazi (2008) has improved latency of the system by 24% over the fixed threshold baseline. Hjalmarsson (2011) also showed significant improvement of reaction time to stimuli with high agreement.

Nevertheless, the problem is not related only to ongoing issues in speech recognition and understanding. For example Ward, Rivera, Ward and Novick (2005) identified turn-taking problems as important shortcomings. The dialogs between a human being and the system are typically straightforward allowing only one speaker at a time. The most common method of recognizing a turn-yield in conversation is waiting for a silent pause longer than a specified threshold. It brings one usability problem: If the user pauses inside an utterance and this pause is longer than the threshold, the user is cut off by the system (Ferrer, Shriberg, & Stolcke, 2002).

Such a simple pause-threshold approach is used infrequently by humans. People tend to use turn exchanges with almost no gap in-between. That was supported by analyses of individual human to human dialog examples (Sacks, Schegloff, & Jefferson, 1978). However, recent work has shown that the 'no-gap in-between dialog turns' dialog is not so common. Heldner and Edlund (2010) explored pauses, gaps and overlaps in three relatively large dialog corpora. Their findings indicate that the timing of turn-taking is not as precise as often claimed. The no-gap-no-overlap model represented less than 1 percent of their data. The gaps preceding the speaker change are long enough to prepare reactive dialog control models, if given published minimal response times for spoken utterances (Shipp, Izdebski, & Morrissey, 1984).

A smooth exchange of turns between dialog partners is supported by various turn-management cues used by the speaker (Duncan, 1972). The turn-taking/yielding cues are sent or received before the moments when the speaker changes. These dialog spots are called Transition Relevance Places (TRPs). This work follows a turn-taking/yielding model which is based on the general turn-taking system defined by Sacks et al. (1978). It has only one party talking at a time. Overlaps (more than one speaker at a time) are less than common and silence gaps are common but brief (Oreström, 1983). The silent party in a dialog uses the speech cues of the speaking one to time properly his/her start of speech. These are called turn-yielding cues (more elaborated in Appendix A – Dialog Turn-Management Cues). For example, one vocal turn-yielding cue is a drop in loudness of speech. There is one example of a dialog sequence in the model (Padilha & Carletta, 2003):

P1: talks low2 low1 TRP DOES NOT TALK;
P2: DOES NOT TALK talks talks;

P1, P2 are dialog participants and low(2-1) is continuous drop in loudness of speech. Research and identification of dialog cues started in 1970s by Duncan's (1972) work.

Table 1 summarizes some of the turn-yielding cues with their strengths and weaknesses. For more extended review of cues see Appendix A. The table offers an overview of many vocal turn-yielding cues thoroughly examined by scientific works.

Unfortunately, concatenative speech synthesis which uses chunks of human records, cannot simply modify these prosodic voice parameters without distorted output sound. The formant speech synthesis does not sound like a human, but the prosodic voice parameters (pitch, speed, etc.) could be simply modified. So there is a possibility to incorporate vocal turn-yielding in formant speech dialog systems.

The main challenge is how to introduce successfully turn-yielding phenomena into a spoken dialog system which uses concatenative speech synthesis to build a system which communicates with the user in a more natural way. One possible solution would be to introduce an ECA in such a dialog system. ECA is able to display/transmit the turn-taking/yielding cues without changing the voice. Having such an ECA-based dialog system would be useful for applications like a voice driven jukebox (Cuřín, Kleindienst, Kunc, & Labský, 2009). There is

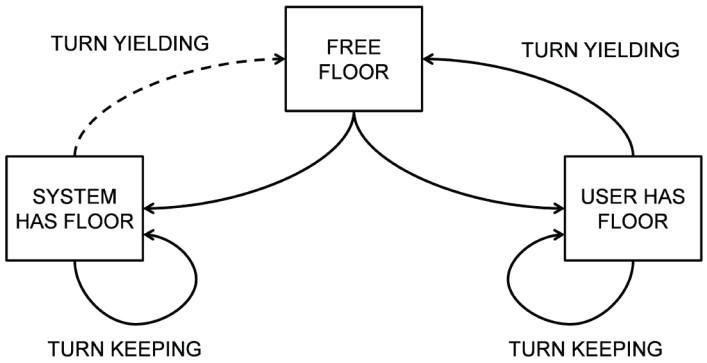
Table 1. Table of strengths and weaknesses of different turn-yielding cues

Category	Turn-Yielding Cue	Strengths	Weaknesses
Vocal	Pitch fall	(Duncan, 1972; Ford & Thompson, 1996) Found in 47% cases before smooth switch (Gravano, 2009) Correctly judged in ~ 60% cases (Hjalmarsson, 2011)	wide variability (Cuttler & Pearson, 1986) Found in 40% cases before smooth switch (Wennerstrom & Siegel, 2003)
	Pitch rise	(Duncan, 1972) Found in 67% cases before smooth switch (Wennerstrom & Siegel, 2003)	Found in 22% cases before smooth switch (Gravano, 2009)
	Higher speaking rate at the end of utterance (reduced lengthening)	Significantly better z-score before smooth switch (Gravano, 2009)	Stressed syllable means slower rate (Duncan, 1972)
	A drop in loudness at the end of utterance	Significantly better z-score before smooth switch (Gravano, 2009)	
	A higher frequency of jitter, shimmer	(Ogden, 2004; Gravano, 2009)	
	Sociocentric sequences	(Duncan, 1972) Judged correctly in ~ 90% cases (Hjalmarsson, 2011)	Okay, yeah sequences overloaded; other cues are needed (Gravano, 2009)
	Syntactical completeness (Textual completeness)	(Duncan, 1972; Sacks, 1974; Schaffer, 1983; Ford & Thompson, 1996; Wennerstrom & Siegel, 2003) necessary cue (Gravano, 2009) Judged correctly in ~ 65% cases (Hjalmarsson, 2011)	
	Stopped movement of head	It's analogy to termination of any hand gestures (Duncan, 1972)	
Visual	Head nod at the end of utterance	Head postural shift – semantic and syntactic boundaries (Kendon, 1972) Head nods as backchannel request (McClave, 2000)	
	Eye-gaze based turn-yielding cue	Multi-party conversations (Gu & Badler, 2006)	Not definite role in two-party conversations (Jokinen, 2010)
Body motoric	The termination of any hand gestures	(Duncan, 1972)	

no known experiment yet which compared the potential of a virtual agent's visual turn-yielding cues to vocal ones. The model of communication is depicted in Figure 1. Taking into account efficiency of communication the simple push-to-talk technology could be very efficient in a speech dialog system as Fernandez, Lucht, Rodriguez and Schlangen (2006) suggest. But, on the other hand, it loses its interactive ability and can be associated with a "vending machine" syndrome (Thórisson, 2002).

We narrowed our area of interest because spoken dialog applications involving only one human user and a computer (two-party) are different from a multi-party approach in terms of turn-taking/yielding cues (smaller space, intimate interaction, importance of some cues) (Jokinen, 2010). Two-party dialogs are a big area of ECA usage as a virtual assistant of companies, for example Rea – a real estate assistant (Cassell, Bickmore, Campbell, Vilhjalmsson, & Yan, 2001). In this work we employ

Figure 1. Diagram of communication. The work is focused on an area of system's turn-yielding (dashed line).



ECA in a spoken dialog system to assess its ability to transmit turn-yielding cues well enough for a human observer to judge them correctly.

First, the study focuses on the number of cues used. That is in line with Duncan's (1972) work. The higher the number of atomic turn-yielding cues at the same point of a conversation the higher the probability of a correct judgment. That reasoning was also supported by experiments (Hjalmarsson, 2011):

H1: Using more turn-yielding cues before a transition relevance place increases the probability of the correct judgment about the next speaker. The turn-yielding cues can be both vocal and visual.

Second, it would be beneficial to learn about the impact of selected visual turn-yielding cues on human judgment in comparison to a selection of the vocal ones. Expectations of a variation in terms of such impact are based on the fact that while one dialog partner listens to the other, his/her hearing system is occupied by hearing the partner's speech. However, the human visual system, which has even more capacity than the hearing system (Card, Moran, & Newell, 1986), is not occupied.

Barkhuysen, Krahmer, and Swerts (2008) studied the relation between auditory and visual turn-taking cues using recorded human conversations. They systematically compared audio-only, vision-only and audio-visual groups of cues. Interestingly, they found out that audio-only cues are less reliable than other groups. So, we can assume that visual turn-yielding cues are more reliable in the process of yielding a turn in a dialog.

Raux and Eskenazi (2008) presented another interesting result. Also an implementation of an end-detection algorithm revealed that prosodic features did not help turn-ending detection once other features (semantic) were included. Duncan (1972) finds the termination of any hand gestures as an important turn-final visual cue. These results and results of other turn-yielding cues (see Table 1) lead us to the following hypothesis:

H2: Visual turn-yielding cues are better than vocal cues in increasing the probability of a correct judgment of who will be the next speaker.

METHOD

To test the delineated hypotheses mentioned above a perceptual experiment is introduced. An investigation of simple and complex turn-yielding signals is done in this experiment where participants watch and listen to videos of conversation between a talking head avatar

as one of the dialog partners (Annie) and a simple vocal dialog partner (David). The talking head is used in order to find out whether this sort of avatar is capable of turn-yielding signaling and whether people can judge this signaling correctly.

The conversations are paused in selected spots and participants try to decide whether the speaker changes or not. The method of analysis of judgments of non-participating dialog listeners is exploited. The methodology of our perceptual experiment is mainly inspired by the previous work and experiments of Hjalmarsson (2009) because of the similarity of investigated issues. Her work was later extended to experiments with synthesized voice. Hjalmarsoon (2011) found out that synthetic voice affects judgments in a similar way as the human voice.

Several turn-yielding cues are employed in the experiment. The selection consists of three vocal turn-yielding cues (pitch fall, higher speaking rate and loudness at the end of utterance) because of their effectiveness (see Table 2) and because they are also re-synthesizable. Two visual turn-yielding cues introduced are:

- movement of head is slowed down before yielding the turn in dialog;
- small head nod at the end of utterance.

They were selected because they can be seen on the face, i.e. the whole body is not needed for their expression. An eye-gaze based turn-yielding cue was not selected. As stated in the literature, eye-gaze based turn-yielding cues are very important in multi-party conversations (Gu & Badler, 2006) but they do not have as definite a role in two-party conversations (Jokinen, 2010).

EXPERIMENTS

Details of the perceptual experiment conducted with the aim to evaluate three vocal and two visual turn-yielding cues are provided in Table 2. The experiment was performed in two parts: first, the main experiment, and then a post-test experiment.

Dialog Data

The dialog part of the experiment was conducted in English because the speech synthesizer (IBM ViaVoice speech synthesizer) used supports English only both in formant and concatenative synthesis. Dialog data is needed to run the experiment successfully. As the speech dialog systems research is a relatively young research area there are no "standard" dialog data which could be used.

There are several possible ways to collect dialog data. First, one can design an experiment (a game) in which participants are forced to talk to each other while playing. Such a situation creates an environment where a natural dialog can be captured. Second, movie scripts serve as another source of dialog data from artificially prepared dialogs by skilled writers.

A part of a theater play script is chosen in this experiment to study turn-yielding cues, using artificially prepared dialogs from a play "The Importance of Being Earnest" by British

Table 2. Table of evaluated turn-yielding cues

Category	Turn-Yielding Cue
Utterance pitch (vocal)	Fall
Utterance final speed (vocal)	Higher
Utterance final loudness (vocal)	Low
Talking head movement	Stopped
Talking head nod	Head nod

author Oscar Wilde. The dialog data source for the second post-test experiment tries to counterbalance this and uses natural dialogs source. Dialog data for the post-test experiment are based on the Map Task Corpus of the University of Edinburgh, the part accessible from Dialogue Diversity Corpus (see http://www-rcf.usc.edu/~billmann/diversity/DDivers-site.htm). These dialogs are not artificially written. They are collected during a map navigation game (Anderson et al.,1992).

Experimental Design

The dialog data should be transformed from text to speech in order to evaluate turn-yielding cues. The speech part of this transformation is done through a speech synthesizer. The synthesizer is able to generate all turn-yielding vocal cues but the pitch fall cue; that only through formant synthesis. Wherever the pitch fall cue is used, the formant synthesis is present. The other dialog sequences use concatenative synthesis. The usage of formant synthesis which is not as good as concatenative synthesis is a possible problematic point. It is further discussed at the end of this work. Visual cues are displayed by a talking head ECA application.

ECA Application

The experiment required an application that is capable of generating ECA sequences and that allows for modification of parameters. The ECAF toolkit is used (Kunc & Kleindienst, 2007). This toolkit is capable of displaying a 3D talking head and as such generates video sequences of an animated talking head. The visual output is lip-synchronized to a synthesized English speech audio signal and saved in a form of MPEG-4 video file.

The talking head continuously moves while speaking according to a pseudo-random movement pattern. The slowed down movement of the head at the end of an utterance is simulated through stop of continuous movement of the talking head. A second visual turn-yielding cue is modeled like a "human head nod". At the end of an utterance the head bends forward for half a second. To avoid possible flaws of experimental data through eye movements the talking head looks synchronously the same way to the user as she moves.

Evaluation Videos

Fifteen dialog scenarios were prepared and generated by the ECAF toolkit. Table 3 contains the list of video, audio and dialog parameters. The directions of dialog turns could not be counterbalanced correctly because David is just a voice and does not have an avatar. He could yield turn just in 'sound only' sequences. However this should not influence the results. Kennedy and Camden (1983) did not found significant gender differences in similar area of interruptions during conversations. Each dialog sequence runs about 20 to 60 seconds. The sentences in dialog sequences were semantically and syntactically complete. The list of utterances just before judgment points is in Appendix B.

Turn-Yielding Cues Parameters

This section lists the main parameters of how the turn-yielding cues are created. Table 4 shows vocal and visual turn-yielding cues creation parameters. Vocal cues are described by relative changes of fundamental frequency F_0. Typically, they include relative speedup at the end of turn, pitch and a volume relative change in dB. The pitch contour pictogram follows ToBi annotation conventions (Beckman & Hirschberg, 1994). Intensity compares sound intensity change at the turn-ending to overall previous intensity in dB. And finally, speed change represents a factor of changed speed at an utterance turn-ending. The length of all vocal turn-yielding cues was about 500 ms.

Visual cues are prepared on the talking head ECA mentioned in the previous section. The talking head movement cue is a very simple one. The talking head constantly moves according to a pseudo-random movement pattern. The cue is displayed as a stop of this movement just before turn-ending. It also has

Table 3. List of evaluated video/dialog parameters in the main experiment

Dialog No.	Turn-Yielding Cues	Direction of Dialog Turn-Yield	Sound-Only
0	Without	David – David	No
1	Without	David – Annie	No
2	Without	David – David	Yes
3	Head movement	Annie – David	No
4	Pitch fall	David – Annie	Yes
5	Final speed	David – Annie	No
6	Final loudness	David – Annie	Yes
7	Head nod	Annie – David	No
8	Final loudness, pitch	David – Annie	Yes
9	Head movement, nod	Annie – David	No
10	Final speed, pitch	Annie – David	Yes
11	Head movement, nod, final speed	Annie – David	No
12	Head movement, nod, final pitch	Annie – David	No
13	Head movement, nod, final loudness, speed, pitch	Annie – David	No
14	Head movement, nod, final loudness, speed	Annie – David	No

Table 4. Audio parameters of turn-yielding cues

Cue	Pitch Contour	F_0 Change	Intensity Change	Speed Change
Pitch fall	%L-L	104-98Hz	0dB	The same
Final loudness	%H-L	182-88Hz	-10dB	The same
Final speed	%H-L	182-88Hz	0dB	1.6x faster
Visual cues	%H-L	210-108Hz	0dB	The same

a length of 500 ms. The second cue, that of head nod, is represented by an animation 500 ms long where the head bends a little bit forward and then returns to the neutral position.

Dialog and Experiment Parameters

Dialog partners are called "Annie" and "David". Therefore, the dialog represents a man-to-woman conversation. Annie's part is played by the talking head avatar and David is invisible (only his speech can be heard). During the preparation of dialog sequences, the man and woman speech synthesis were used to differentiate the roles of Annie and David. The dialogs were prepared to be interrupted during or at the end of Annie's or David's utterance. In this TRP point either Annie or David holds the turn or yields it to the partner.

Figure 2 depicts arrangement of the experiment. The test was driven by Adobe Flash web application (for screenshots see Figure 3) which played each dialog video to the participant and asked questions. The precise procedure of the experiment is given in the next section. Each

Figure 2. Experiment set up. Annie is the ECA with a synthesized voice and she talks to David (he is just synthesized voice). The observer follows a dialog sequence between Annie and David and this dialog sequence is interrupted. The observer judges who will be the next speaker (Annie or David) after the dialog sequence is interrupted.

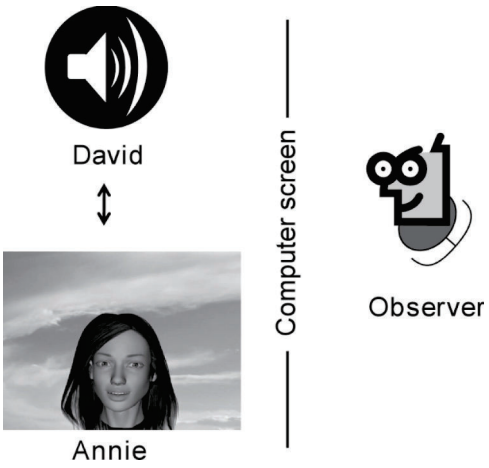
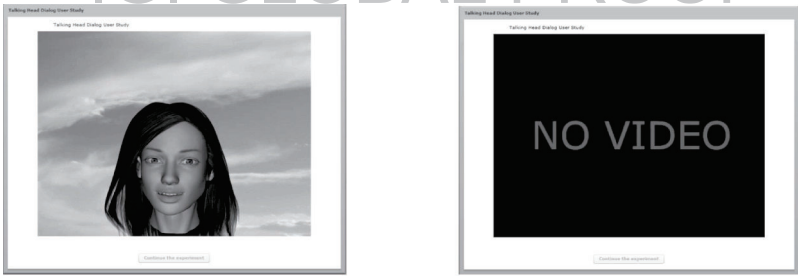


Figure 3. Experiment web application screenshots: a) Talking head dialog video, b) Sound-only dialog sequence example, c) Question: "Who will be the next speaker"?



(a) (b)

(c)

of the videos is 640 x 480 pixels (190 x 142.5 mm) large and participants saw the videos in this size (it was not possible to scale down or up).

Participants

In total 40 participants took part in our comparison experiment, the main part. A homogeneous group of participants was needed in terms of knowledge of English. The participants were both males (50%) and females (50%) – mean age 24.70 (*SD* 2.20), and they were university students of computer science, philosophy and political science with a fairly good knowledge of English. The participants had minimally four one-semester courses of English at the university that ended with a final exam. Dialog sequences for the experiment were chosen so that the participants would understand them. Participants had Czech cultural background. Before the experiment, each of the participants was instructed verbally on how to proceed with the experiment. The whole experiment was anonymous. Some participants were motivated by receiving a small amount of points in their human computer interaction course while the others wanted to help our department. The next section describes the experimental routine properly.

Experiment Procedure

The experiment was conducted remotely. Every participant observed the dialog sequences in their computer in a different but quiet environment at home in their spare time. "How to proceed the experiment" instructions were given verbally in the teaching lab. The possibility that one person would perform the experiment multiple times was excluded by the usage of faculty login name.

The method of non-participating judgments was followed. A Flash web application was developed, mainly because of the ability of video playback. The first page of the web application contained the login field (to exclude multiple participation) and written instructions. After login, the participant navigated to the first video dialog. The participant should see the screen with video (Figure 3a) and hear the sound, in case the sequence was recorded with a talking head. In the case of sound-only dialog, depicted in Figure 3b, the participant heard only the sound of the dialog. The "NO VIDEO" (Figure 3b) screen was shown to the participants before the experiment. This sign should not affect the results because it is known in advance. Finally, the particular turn-yielding cues are presented (or not) and the video suddenly ends. The participant is asked the question: "Who will be the next speaker? Annie or David?" (see Figure 3c). The possible answers are: "Annie", "David", or "I was interrupted". The last option allows the participant, in case he/she was interrupted during the video observation (for example, by a phone call, another person, etc.), to repeat this particular dialog case. After he/she answers the question he/she can proceed to the next dialog sequence. The "I was interrupted" allows a participant to repeat the video. This could invalidate the whole experiment, but nobody used this "I was interrupted" button.

Each participant judged 15 spots in dialogs (one judgment spot at the end of each dialog sequence). During the test, it was possible to suspend the procedure between dialog sequences and finish it later. Random order of dialog sequences was generated for each participant to counterbalance the learning curve effect. Correct answers were not revealed to participants during the experiment procedure. They were revealed to them after all participants successfully conducted the experiment.

Experiment Evaluation and Discussion

In total 40 participants completed our experiment. Participants answered a total of 600 questions (mentioned above). The application recorded one answer for each question. The

possible answers for each question are: "Annie" or "David". The results were converted to the graph in Figure 4. The horizontal axis shows all videos and the vertical axis represents the percentage of correct answers.

The results were analyzed for gender differences. Chi-square test with Yates' continuity correction was applied to each dialog sequence. The tests revealed that there are no significant differences between the genders (all p-values were > 0.1). The largest difference was in the case of dialog sequence 5 (final speed, $p = 0.11$), nevertheless, this difference is still not significant.

It can be seen that only in dialog sequences 0, 1, 4 and 8 majority of participants judged incorrectly. Sequences in the right side of the graph show very good results. According to Table 2, these dialog sequences used mostly combinations of cues. To evaluate both hypotheses the experiment data were statistically analyzed.

Hypothesis H1

As said at the beginning of Section Experiment, hypothesis H1 examines usage of turn-yielding cue combinations in one TRP. The decision process after each of the dialog sequences is in fact a Bernoulli trial because of the two options as to who will be the next speaker (Papoulis, 1984). Having the number of used turn-yield cues in each dialog sequence, logistic regression statistics could be used. We used binary ordinal logistic regression model and all statistical computations were done in R software (R Development Core Team, 2011).

The results of logistic regression show that there is a relation between the number of used turn-yielding cues and correctness of participant judgment. Table 5 shows coefficients on how the correctness of judgment changes when a particular number of turn-yielding cues is used (odds ratios). The counts of turn-yielding cues are split to three groups to lower degrees of freedom and to provide better fit of a final model. Three results are statistically significant and they improve the correctness of judgment substantially. It supports the hypothesis H1.

Likelihood ratio test $\chi 2$ of 12.84 with 3 degrees of freedom and $p = 0.00$ show that our model fits significantly better than empty (null) model. We don't report Cox & Snell R^2, Nagelkerke R^2 or other R^2 measures because they do not assess goodness-of-fit (Hosmer & Lemeshow, 2000). R^2 measures are based on comparisons of predicted values from the fitted model to intercept or null model only; however they may be helpful in the model building state for evaluating competing models (Hosmer & Lemeshow, 2000). Measures of fit are based

Figure 4. Judgment correctness of participants in the main experiment. The results of male and female participants were expressed separately to reflect possible differences between genders. The most correct answers were from dialog sequence no. 11. A combination of three turn-yielding cues (head movement, head nod and final speed) was used there. See Table 2 for other dialog sequences description.

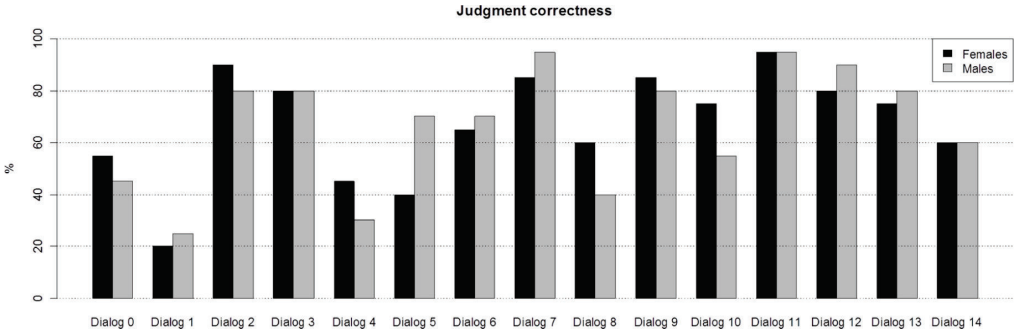


Table 5. Odds ratios of how correct judgment improves with more cues used

Number of Cues	Odds Ratios	*p*-Value
1	0.51	$p = 0.06$
2, 4)	0.98	$p = 0.00$
4, 5	0.64	$p = 0.05$

on a comparison of observed to predicted values from the fitted model. The correct classification table is superseded by the following goodness-of-fit test.

Goodness-of-Fit Test

The Hosmer-Le Cessie statistic test is a measure of lack of logistic regression model fit. This statistic test is based on the Hosmer-Lemeshow test, which divides the data into groups of equal size, and then compares the observed to expected number of positive responses and performs a $\chi 2$ test. The Hosmer-Le Cessie test solves some insufficiencies of the original test (Hosmer, Hosmer, Cessie, & Lemeshow, 1997). The test was computed for the hypothesis H1 data. Unweighted sum-of-squares is 118.95, expected mean value is 118.95, variance <0.00001 and p = 1.00. Very large value of sum-of-squares would indicate lack of fit, but this value seems to be low enough, so the logistic regression model doesn't show a lack of fit. The *p*-value is high enough not to reject the null hypothesis that the data follow the logistic regression model (in Table 4).

Hypothesis H2

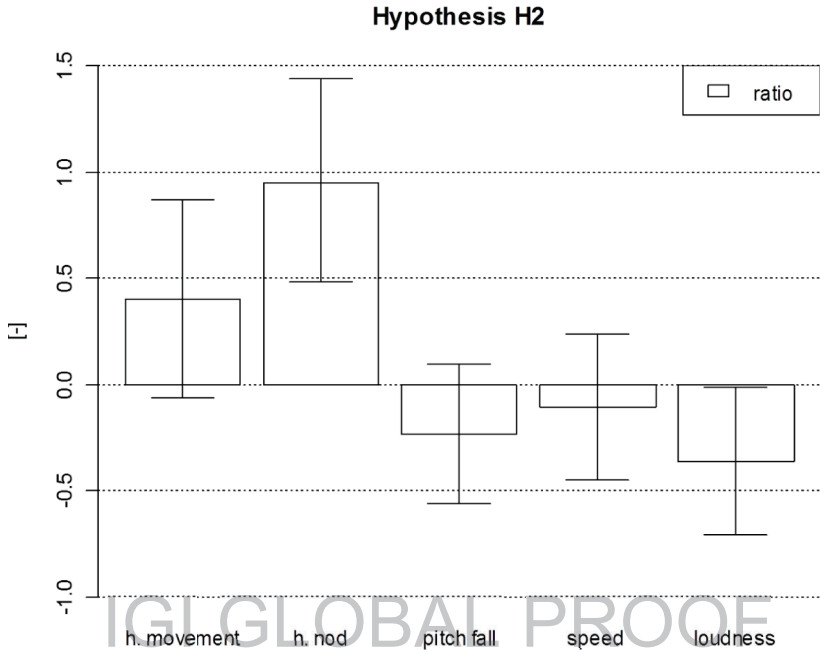
The hypothesis H2 says that visual cues are better than vocal ones. Logistic regression can also validate this hypothesis when the experiment results are separated to individual variables. Figure 5 shows the results of logistic regression and 90% confidence intervals. The pitch fall cue ($p = 0.24$), the head movement cue ($p = 0.15$) and the final speed ($p = 0.61$) results are not significant.

Figure 4 can be viewed that the hypothesis H2 is valid. The potential of visual cues appears to be better. However, the results of this experiment are misleading because the dialog sequences contain a mixture of visual and vocal cues. Better insight can be provided by another experiment which will have visual and vocal cues separated. Likelihood ratio test $\chi 2$ of 45.38 with 5 degrees of freedom and $p = 0.00$ shows that our model fits significantly better than empty (null) model. The Hosmer-Le Cessie goodness of fit test turned out to be as following: Unweighted sum-of-squares is 112.49, expected mean value is 112.23, variance 0.21 and $p = 0.19$. The *p*-value is reasonably high and the model fits well. The experiment in the next section can bring clearer view on hypothesis H2.

POST-TEST EXPERIMENT

The post-test experiment focuses on hypothesis H2 and brings a less distorted view on this hypothesis in that the visual and vocal cues are not mixed in dialog sequences. The conditions and execution of the experiment is exactly the same as during the experiment in Section Experiment. Although the number of participants is smaller than in the previous experiment, it is sufficient for us to show validity of the hypothesis H2 as the findings are statistically significant. The main difference is the usage of Map Task dialog source mentioned in Section Dialog Data which brings real dialog utterances in contrast to Oscar Wilde's play.

Figure 5. Graph of the logistic regression results for hypothesis H2. It shows ratios how probability of correct judgment rises or falls using particular turn-yielding cue. Visual turn-yielding cues (head movement and head nod) raise the probability. Vice versa the vocal turn-yielding cues decrease or leave the same probability of correct judgment.



The video/audio dialogs were created in the same way with Annie's and David's speech synthesized. The parameters and dialog setup could be found in Table 6. Ten dialog sequences were prepared and used the same dialog in all sequences. This time when the vocal cues are utilized in a dialog, only the audio is evaluated. Naturally, visual cues required the video dialogs. Concatenative synthesis was used for all dialog sequences and pitch fall cue was simulated by Praat tool (http://www.fon.hum.uva.nl/praat/).

As in the previous experiment, sentences in dialog sequences were semantically and syntactically complete. List of the utterances preceding the judgment point are in Appendix B.

This time the experiment was conducted with a total of 31 participants. The participants were males (65%) and females (35%) – mean age 26.60 (*SD* 3.60). Not a single participant took part in the previous experiment. This group of participants was more heterogeneous than the previous one. There were not only students involved. Participants had Czech cultural background. We wanted to compensate for the previous homogenous group of students and examine the results in a different group of participants. The participants went through prepared dialog sequences.

The results of judgment correctness are summarized by the graph in Figure 6. The results of the turn-yielding cues of head movement, head nod and sound loudness are statistically significant. The results of final speed and falling pitch are not statistically significant. See Table 7 for particular odds ratios. As it can be seen, the odds ratio of vocal cue loudness improved in comparison to the previous experiment, but the final pitch still has a negative effect on judgment correctness (in this experiment not statistically significant). The improvement of loudness vocal cue could be realized by different dialog

Table 6. List of post-test evaluation video/dialog parameters

Dialog No.	Turn-Yielding Cues	Direction of Dialog Turn-Yield	Sound-Only
0	Without	David – David	No
1	Without	Annie – Annie	Yes
2	Head movement	Annie – David	No
3	Final speed	Annie – David	Yes
4	Head nod	Annie – David	No
5	Final loudness	Annie – David	Yes
6	Head movement, nod	Annie – David	No
7	Pitch fall	Annie – David	Yes
8	Without	Annie – Annie	No
9	Final loudness, speed, pitch fall	Annie – David	Yes

Figure 6. Judgment correctness of participants in the post-test experiment. See Table 5 for dialog sequences description.

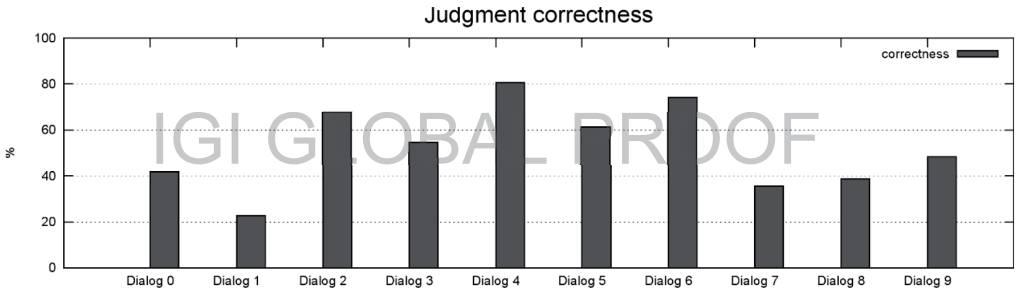


Table 7. Odds ratios of how correct judgment improves with more cues used. Statistically significant results are in bold.

Turn-Yielding Cue	Odds Ratios	*p*-Value
Head movement	0.71	$p = 0.04$
Head nod	1.28	$p = 0.00$
Pitch fall	-0.46	$p = 0.17$
Final speed	0.36	$p = 0.28$
Loudness	0.61	$p = 0.05$

sequences and by the usage of qualitatively better concatenative synthesis in comparison to the formant synthesis. Likelihood ratio test $\chi 2$ of 29.72 with 5 degrees of freedom and $p = 0.00$ show that our model fits significantly better than empty (null) model. The Hosmer-Le Cessie goodness of fit test was also done and the values are the following: unweighted sum-of-squares is 71.51, expected mean value is 71.87, variance 0.15 and $p = 0.06$. There does not seem to be a

lack of fit. The statistical non-significance of two vocal turn-yielding cues seems to also be caused by very indecisive results for these cues. But the overall judgment correctness of our selection of visual turn-yielding cues supports the hypothesis H2 that they are more reliable than the vocal ones.

CONCLUSION

In this paper we examined the use of turn-yielding elements in speech dialog systems. Employment of selected visual turn-yielding cues was studied where possible and where the speech synthesis did not allow for modifying prosodic parameters for vocal turn-yielding cues in real time. The main motivation was to build a more natural embodied conversational agent (ECA).

Two non-participating judgment experiments were performed and five visual and vocal turn-yielding cues were compared. A total of 71 participants accomplished both experiments. Each participant judged 15 dialog sequences in the first experiment or 10 dialog sequences in the second one. Findings from the first main experiment suggest that the hypothesis H1 (Using more turn-yielding cues before a transition relevance place increases the probability of correct judgment of the next speaker) is valid. Possible gender differences in the results were also analyzed. The results of statistical tests show that there are no significant differences in turn-yielding cues perception between male and female participants.

The second hypothesis (H2) that visual cues are more reliable than vocal turn-yielding cues was also validated by the first experiment. However the pitch fall cue, the head movement cue and the final speed cue results are not significant. That could be caused by more indecisive judgment results. Not to be misled by the mixture of visual and vocal turn-yielding cues, the post-test experiment was prepared. The post-test experiment validated the second hypothesis. It turned out that our selected visual turn-yielding cues were more reliable than the vocal ones (H2) in the post-test experiment. Finally, the two experiments suggest that usage of selected visual turn-yielding cues has a positive impact on dialog turn management (better turn-ending estimates) in the area of two-party dialogs.

The findings concerning vocal turn-yielding cues are a little bit surprising. The selected vocal turn-yielding cues seem to have little impact on correct turn-ending judgment. Furthermore, pitch fall cue had a negative effect on this judgment in the first experiment as well as in the second experiment, however, the findings were not statistically significant. These results are in contradiction to previous studies considering vocal turn-yielding cues (e.g. Gravano, 2009; Hjalmarsson, 2009; Hjalmarsson, 2011). One possible explanation could be the length of audio turn-yielding stimuli. According to previous studies of Barkhuysen et al. (2008) audio-only features are better classified when they are longer. Or, in the case of the first main experiment, this could be caused by a formant synthesis quality.

Comparison of vocal and visual cues may seem a little bit "unfair" because the voice in the experiment had several functions (e.g. turn-taking and delivering intelligible speech). But the talking head had also several functions like the voice, turn-taking and delivering intelligible speech in form of face animation.

Although the post-test experiment tried to solve some shortcomings/weaknesses of the first experiment (e.g. vocal and visual cues mixed in dialogs, formant speech synthesis, etc.), some confounding points could still exist. Those should be addressed in future work. For example, the experiments were not fully gender or dialog modality counterbalanced because of non-existent male avatar. We used human-like female avatar only. Therefore, the results cannot be generalized and are limited to the conditions of the two experiments.

Still, the results of the experiments show clearly the advantage of ECA usage in speech dialog systems. The spoken dialog system architects have the ability to include turn-yielding

cues of ECA and use state-of-the-art concatenative synthesis together in their systems. This could make the system more natural and improve its interactivity. Efficiency of interaction seems to be very good even in push-to-talk systems (Fernandez, Lucht, Rodriguez, & Schlangen, 2006).

In future, we would like to continue investigating turn-yielding cues and add turn-taking cues, as well as experiment with the whole body gestures. Some other turn-yielding cues effects were already explored by Hjalmarsson (2011), e.g. cue phrase – response eliciting cue had good turn-yielding results. Also, focus on the whole body provides us with even more turn-taking/yielding cues such as body posture, hand posture, etc. It would also be interesting to explore how judgment results develop when the ECA uses a wider range of gestures, not only a speech animation and turn-yielding cues. Such gestures could soften movement differences in turn-endings and the users could be more confused by these gestures while detecting turn-endings. In addition, it might be interesting to examine whether the results stay the same when just a basic head or a cartoon head is used as an avatar.

ACKNOWLEDGMENT

This research has been partially supported by the MSMT under the research program LC-06008 (Center for Computer Graphics). This research has been also partially supported by the MSMT under the research program MSM 6840770014. This paper has been also partially supported by the EC funded project VERITAS, contract number FP7 247765. Thanks to the participants of experiments.

REFERENCES

Anderson, A. H., Bader, M., Bard, E. G., Boyle, H., Doherty, G. M., & Garrod, S. C. et al. (1992). The HCRC map task corpus. *Language and Speech*, *34*(4), 351–366.

Barkhuysen, P., Krahmer, E., & Swerts, M. (2008). The interplay between auditory and visual cues for end-of-utterance detection. *The Journal of the Acoustical Society of America*, *123*(1), 354–365. doi:10.1121/1.2816561 PMID:18177165.

Beckman, M., & Hirschberg, J. (1994). *The ToBI annotation conventions*. Ohio State University.

Card, S., Moran, T., & Newell, A. (1986). *The model human processor - An engineering model of human performance*. John Wiley and Sons.

Carlson, R., & Hirschberg, J. (2009). *Cross-cultural perception of discourse phenomena. Proc. of Interspeech* (pp. 1723–1726). ISCA.

Cassel, J. (2000). *Embodied conversational agents*. MIT Press.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., & Yan, H. (2001). More than just a pretty face: Conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, *14*(1-2), 55–64. doi:10.1016/S0950-7051(00)00102-7.

Cuřín, J., Kleindienst, J., Kunc, L., & Labský, M. (2009). Voice-driven Jukebox with ECA interface. In *Proceedings of the 13th International Conference Speech and Computer* (pp. 146-151).

Cuttler, A., & Pearson, M. (1986). On the analysis of prosodic turn-taking cues. In C. Johns-Lewis (Ed.), *Intonation and discourse* (pp. 139–155). Cambridge, UK: Cambridge University Press.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2). doi:10.1037/h0033031.

Duncan, S., & Fiske, D. (1977). *Face-to-face interaction: Research, methods and theory*. Lawrence Erlbaum Associates.

Edlund, J., Heldner, M., & Gustafson, J. (2005). Utterance segmentation and turn-taking in spoken dialogue systems. *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen,* 576-587.

Fernandez, R., Lucht, T., Rodriguez, K., & Schlangen, D. (2006). Interaction in task-oriented human-human dialogue: The effects of different turn-taking policies. *Spoken Language Technology Workshop* (pp. 206-209). IEEE.

Ferrer, L., Shriberg, E., & Stolcke, A. (2002). is the speaker done yet? In *Proceedings of the ICSLP* (pp. 2061-2064).

Ford, C. E., & Thompson, S. A. (2010). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the projection of turn completion. *Pragmatics*, *4*(1), 31–62.

Gorin, A. L., Riccardi, G., & Wright, J. H. (1997). How may I help you? *Speech Communication*, *23*(1-2), 113–127. doi:10.1016/S0167-6393(97)00040-X.

Gravano, A. (2009). *Turn-taking and affirmative cue words in task-oriented dialogue*. Columbia University.

Gu, E., & Badler, N. (2006). *Visual attention and eye gaze during multiparty conversations with distractions. Intelligent Virtual Agents* (pp. 193–204). Springer. doi:10.1007/11821830_16.

Hargie, O., & Dickson, D. (2004). *Skilled interpersonal communication: Research, theory, and practice*. Psychology Press.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568. doi:10.1016/j.wocn.2010.08.002.

Heylen, D. K. (2009). Understanding speaker-listener interaction. In *Proceedings of the Interspeech Conference* (pp. 2151-2154). International Speech Communication Association.

Hjalmarsson, A. (2009). On cue - additive effects of turn-regulating phenomena in dialogue. In *Proceedings of the Diaholmia.*

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, *53*, 23–35. doi:10.1016/j.specom.2010.08.003.

Hosmer, D., Hosmer, T., Cessie, S. L., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, *16*(9), 965–980. doi:10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O PMID:9160492.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. Wiley. doi:10.1002/0471722146.

Jokinen, K. (2010). Non-verbal signals for turn-taking and feedback. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC)* (pp. 2961-2967).

Jokinen, K., Nishida, M., & Yamamoto, S. (2009). Eye-gaze experiments for conversation monitoring. In *Proceedings of the 3rd International Universal Communication Symposium* (pp. 303-308). ACM.

Jonsdottir, G., & Thórisson, K. (2009). *Teaching computers to conduct spoken interviews: Breaking the realtime barrier with learning* (pp. 446–459). Springer. doi:10.1007/978-3-642-04380-2_49.

Jonsdottir, G., Thórisson, K., & Nivel, E. (2008). *Learning smooth, human-like turntaking in realtime dialogue. Intelligent Virtual Agents* (pp. 162–175). Springer. doi:10.1007/978-3-540-85483-8_17.

Kendon, A. (1972). Some relationship between body motion and speech. In A. Siegman, & B. Pope (Eds.), *Studies in dyadic communication* (pp. 177–210). New York, NY: Pergamon Press.

Kok, I. d., & Heylen, D. (2009). Multimodal end-of-turn prediction in multi-party meetings. In *Proceedings of the International Conference on Multimodal interfaces* (pp. 91-98). ACM.

Kok, I. D., & Heylen, D. (2011). The MultiLis corpus-dealing with individual differences in nonverbal listening behavior. *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, 362-375.

Kronild, F. (2006). Turn-taking for artificial conversational agents. In *Proceedings of the Cooperative Information Agents* (pp. 81-95). Springer.

Kunc, L., & Kleindienst, J. (2007). ECAF: Authoring language for embodied conversational agents. In *Proceedings of the TSD* (pp. 206-213). Springer.

Lai, J., & Yankelovich, N. (2008). Conversational speech interfaces. In *The human-computer interaction handbook* (2nd ed., pp. 381–392). Lawrence Erlbaum Associates Inc..

Maat, M. t., & Heylen, D. (2009). *Turn management or impression management? Intelligent Virtual Agents* (pp. 467–473). Springer. doi:10.1007/978-3-642-04380-2_51.

Maat, M. t., Truong, K., & Heylen, D. (2010). *How turn-taking strategies influence users' impressions of an agent. Intelligent Virtual Agents* (pp. 441–453). Springer. doi:10.1007/978-3-642-15892-6_48.

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, *32*(7), 855–877. doi:10.1016/S0378-2166(99)00079-X.

Mortensen, C. D. (2007). *Communication theory*. Transaction Publishers.

Noguchi, H., & Den, Y. (1998). Prosody-based detection of the context of backchannel responses. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, (pp. 487-490).

Ogden, R. (2004). Non-modal voice quality and turn-taking in Finnish. In E. Couper-Kuhlen, & C. E. Ford (Eds.), *Sound patterns in interaction: Cross-linguistic studies from conversation* (pp. 29–62). Amsterdam, Netherlands: John Benjamins Publishing Co..

Oliveira, M., & Freitas, T. (2008). Intonation as a cue to turn management in telephone and face-to-face interactions. In *Proceedings of the Speech Prosody* (pp. 485-488). ISCA.

Oreström, B. (1983). *Turn-taking in English conversation*. Krieger Publication Co..

Padilha, E., & Carletta, J. (2003). Nonverbal behaviours improving a simulation of small group discussion. In *Proceedings of the 1st Nordic Symp. on Multimodal Communication* (pp. 93-105).

Papoulis, A. (1984). *Probability, random variables, and stocastic processes* (2nd ed.). McGraw-Hill.

Pieraccini, R., Suendermann, D., Dayanidhi, K., & Liscombe, J. (2009). Are we there yet? Research in commercial spoken dialog systems. In *Proceedings of the TSD. 5729* (pp. 3-13). Springer.

Poppe, R., Truong, K., Reidsma, D., & Heylen, D. (2010). *Backchannel strategies for artificial listeners. Intelligent Virtual Agents* (pp. 146–158). Springer. doi:10.1007/978-3-642-15892-6_16.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raux, A., & Eskenazi, M. (2008). Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue* (pp. 1-10). ACL.

Raux, A., & Eskenazi, M. (2009). *A finite-state turn-taking model for spoken dialog systems* (pp. 629–637). Association for Computational Linguistics. doi:10.3115/1620754.1620846.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turntaking for conversation. In *Studies in the Organization of Conversational Interaction* (pp. 7–55). Academic Press.

Schaffer, D. (1983). The role of intonation as cue to turn taking in conversation. *Journal of Phonetics*, *11*, 243–257.

Schlangen, D. (2006). From reaction to prediction: Experiments with computational models of turn-taking. In *Proceedings of the Interspeech.* ICSA.

Shipp, T., Izdebski, K., & Morrissey, P. (1984). Physiologic stages of vocal reaction time. *Journal of Speech and Hearing Research*, *27*, 173–178. PMID:6330455.

Thórisson, K. (2002). Natural turn-taking needs no manual: Computational theory and model, from perception to action. *Multimodality in language and speech systems*, 173-207.

Ward, N., & Bayyari, Y. A. (2010). American and Arab perceptions of an Arabic turn-taking cue. *Journal of Cross-Cultural Psychology*, *41*(2), 270–275. doi:10.1177/0022022109354644.

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, *32*(8), 1177–1207. doi:10.1016/S0378-2166(99)00109-5.

Ward, N. G., Rivera, A. G., Ward, K., & Novick, D. G. (2005). Root causes of lost time and user stress in a simple dialog system. In *Proceedings of the Interspeech.*

Wennerstrom, A., & Siegel, A. F. (2003). Keeping the floor in multiparty conversations: Intonation, syntax, and pause. *Discourse Processes*, *36*(2), 77–107. doi:10.1207/S15326950DP3602_1.

Wiemann, M., & Knapp, M. L. (1975). Turn-taking in conversations. *The Journal of Communication*, *25*, 75–92. doi:10.1111/j.1460-2466.1975.tb00582.x.

Yee, N., Bailenson, J. N., & Rickertsen, K. (2007). A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In *Proceedings of the CHI* (pp. 1-10). ACM.

Yngve, V. (1970). On getting a word in edgewise. In *Proceedings of the Sixth Regional Meeting Chicago Linguistic Society* (pp. 567-578).

Ladislav Kunc is a PhD student at the Department of Computer Graphics and Interaction at Czech Technical University in Prague, Faculty of Electrical Engineering. Since 2006 he has worked at the same time as a researcher in IBM Prague R&D Lab. His main area of interest is human computer interaction with embodied conversational agents. Especially, he is interested in the design of multimodal applications and speech-based dialog systems. Recently, he has extended his focus on the dialog systems for car environments. He is an author or co-author of several publications in the fields of human computer interaction and embodied conversational agents.

Zdenek Mikovec works at CTU FEE in the Department of Computer Graphics and Interaction as a researcher and teacher and is in charge of the university's usability lab. He has co-lead many international research projects funded by the European Union (Mummy, ELU, i2home, VitalMind, AEGIS, ACCESSIBLE). His main focus is on the interaction needs of disabled people, mainly with visual impairments, and interaction in special environments (interactive TV, mobile). He is an author or a co-author of more than twenty publications on various international events in the field of HCI. Web: http://dcgi.felk.cvut.cz/en/members/xmikovec.

Pavel Slavik is a Full Professor at Department of Computer Graphics and Interaction (Czech Technical University in Prague). His professional interests include computer graphics (namely Information visualization) and Human-Computer Interaction (usability, accessibility, design of special user interfaces). He is author or coauthor of more than 200 papers published in journals and in international conferences. He served as an International Programme Committee member in many conferences like Eurographics, SIGCHI, ISVC and many others. He and his team participated in many national and international projects (especially in EC funded projects like IST 5th, 6th and 7th Framework Programme). He was also organizer or co organizer of several conferences and workshops in fields of his interests.

Appendix F

Understanding Formal Description of Pitch–Based Input

Polacek O., Mikovec Z.: Understanding Formal Description of Pitch–Based Input. In Proc. of In Human–Centred Software Engineering, Third International Conference, HCSE2010. Heidelberg: Springer, 2010, vol. 6409, p. 190-197. ISSN 0302-9743.ISBN 978-3-642-16487-3.

Understanding Formal Description of Pitch-Based Input

Ondřej Poláček and Zdeněk Míkovec

Faculty of Electrical Engineering, Czech Technical University in Prague,
Karlovo nam. 13, 12135 Prague 2, Czech Republic
{polacond,xmikovec}@fel.cvut.cz

Abstract. The pitch-based input (humming, whistling, singing) in acoustic modality has already been studied in several projects. There is also a formal description of the pitch-based input which can be used by designers to define user control of an application. However, as we discuss in this paper, the formal description can contain semantic errors. The aim of this paper is to validate the formal description with designers. We present a tool that is capable of visualizing vocal commands and detecting semantic errors automatically. We have conducted a user study that brings preliminary results on comprehension of the formal description by designers and ability to identify and remove syntactic errors.

Keywords: Non-verbal Vocal Interaction; Vocal Gesture; Formal Description; User Study.

1 Introduction

The *Non-Verbal Vocal Interaction* (NVVI) can be described as a method of interaction, in which sounds, other than speech, are produced. There are several approaches described in the literature which include using pitch of a tone, length of a tone, volume, or vowels in order to control the user interfaces. The NVVI is an interaction method that has already received a significant focus within the research community. It has been used as an input modality for people with motor disabilities [7][3] as well as voice training tool [2]. It is a method that shares some similarities with *Automatic Speech Recognition* (ASR). However, when comparing both interaction styles, several differences are revealed. Several reports, including mouse emulation [1] or controlling real-time games [7], suggest that NVVI is better fitted to continuous control rather than ASR. NVVI is cross-cultural and language independent [8]. Unlike ASR, NVVI generally employs simple signal processing methods [3]. Due to NVVIs limited expressive capabilities, ASR is better at triggering commands, macros or shortcuts. NVVI should be considered as a complement to ASR rather than replacement.

To design an application controlled by speech a set of word patterns or grammar must be defined. This grammar will then allow the ASR to recognize a range of expected words used in utterances. Likewise, a designer can also use a similar formal method for pitch-based NVVI.

R. Bernhaupt et al. (Eds.): HCSE 2010, LNCS 6409, pp. 190–197, 2010.

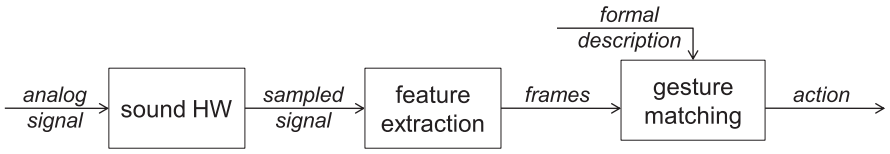


Fig. 1. NVVI signal processing pipeline

The signal processing pipeline for most pitch-based NVVI systems is depicted in Figure 1. Pitch is extracted from the sampled signal in a short discrete periods of time called frames. The typical duration of one frame is approximately 20 ms. The formal description of the NVVI and a stream of frames are then matched together, followed by generation of an appropriate action.

2 Formal Description

When designing a set of voice gestures, the designer must describe an ideal pitch profile for each gesture. These ideal pitch profiles are then referred to as *gesture templates* and they are usually represented in graphic form as shown in Figure 2. However, the users are unable to produce an ideal pitch profile. The interpretation of gesture templates by the user is referred to as *gesture instances*. An example of the relationship between a gesture template and its instances is depicted in Figure 2. Note that slightly different instances share the same semantics defined by the gesture template which is in this case an increasing tone. Once gesture templates are designed in a graphic form, they can be described by a *Voice Gesture Template* (VGT) expressions. Design of VGT expression is described in detail in [5]. These expressions are similar to regular expressions. They have two terminal symbols p and s that correspond to pitch and silence. They also use an operator * for repetition and operator | for the choice. However, there are several symbols with different meanings, for example brackets [ ] which are used for more sophisticated conditions and brackets <> which are used for output definitions to trigger an action. The use of VGT expressions is illustrated in Figure 3. The gesture template depicted in Figure 3 describes instances which start under midi note 60 and increase in pitch to more than 4 midi notes. Midi notes [4] are numerical representations of traditional notes in western music notation, for example, midi note number 60

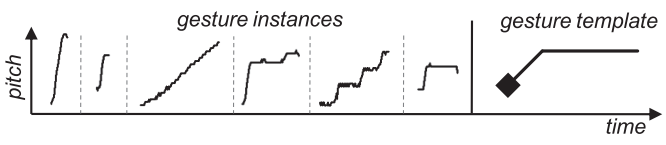


Fig. 2. Relationship between a gesture template and its instances

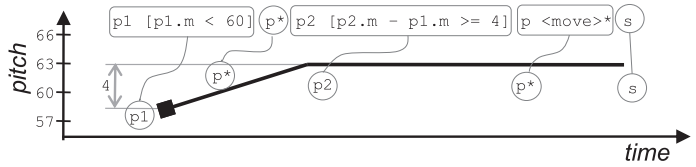


Fig. 3. VGT expression and its graphical representation of gesture template

corresponds to c'. The Figure 3 also illustrates the relation of VGT expression and the graphic representation of the gesture template. This process can be divided into four parts:

1. In the first part, the frame `p1` is matched to the expression when its pitch is under midi note 60. This is ensured by the condition `[p1.m < 60]` where the attribute `.m` is a midi note value of the frame `p1`;
2. Then all pitch frames `p*` are matched until the difference between the pitch of a current frame and the frame `p1` is higher than or equal to 4 midi notes (frame `p2`). This is ensured by the condition `[p2.m - p1.m >= 4]`;
3. After satisfying the condition in the 2nd step, all pitch frames `p <move>*` are matched and the output symbol `move` is triggered with each frame;
4. The processing of the template is completed, when a silent frame `s` is matched.

3 Semantic Errors

Semantic information, that describes pitch profiles of gesture templates, is encoded by a VGT expression. However, the description of gesture templates may be affected by semantic errors which cannot be detected while parsing the expression. A semantic error can also appear in a VGT expression when a new gesture template is added to the expression. The expression must be checked by tedious experimenting that involves user input to see if all templates are recognized correctly. Our research has identified two frequent types of semantic errors which cause improper behavior in gesture recognition – *ambiguous* and *unreachable* templates.

Two gesture templates are ambiguous if there is at least one gesture instance that satisfies both templates. The reason this error frequently occurs is due to an imprecise template description. In a real application there is typically a large number of instances fulfilling the condition of ambiguity. This semantic error is typically demonstrated by the generation of two or more output symbols in one frame.

The gesture template is unreachable when there is no instance matching the template. This can, for example, be caused by a condition that is always false, the template does not take into account human capabilities, or there is another gesture template that prevents the unreachable template from matching instances.

3.1 Semantic Error Detection

Detection of semantic errors, which are described above, requires analysis of gesture instances that can be generated by a VGT expression. We have implemented a tool which is capable of displaying possible gesture instances and automatically identifying semantic errors. It also allows deeper understanding of matching an instance to an expression by tracking its pitch profile. After generating all possible instances that match the expression, the tool checks if each instance belongs to just one template (ambiguity condition) and if each template has at least one instance (unreachability condition).

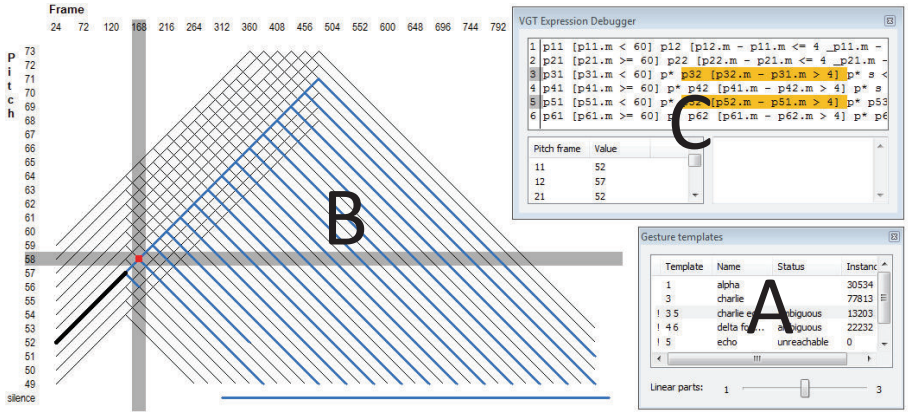


Fig. 4. Tool for vocal gestures visualization and semantic error detection

The user interface of our tool is depicted in Figure 4. Part A of the Figure shows a dialog which contains a list of templates and number of instances. The dialog shows semantic errors within the VGT expression by displaying both ambiguous and unreachable templates (see the *Status* column). The user can display instances by selecting an appropriate row. When selecting a row with ambiguous templates, instances cause the ambiguity are displayed in part B.

Gesture instances are shown in the part B of Figure 4. The horizontal axis represents frames converted into timestamps in milliseconds and the vertical axis represents pitch using midi note numbers [4] starting with silence at the bottom. The black lines represent the generated gestures. When there are a lot of instances and their typical pitch profile is not visible, the user can display these instances and track them from the beginning to the end. When tracking an instance, the corresponding position of a VGT expression is highlighted in the VGT Expression Debugger (dialog in part C). Horizontal and vertical bars represent the current position, the bold line represents the part of an instance that has been already tracked and the blue lines show the further extending of a current instance.

The VGT expression is shown in dialog C. The current position of tracked instance is highlighted directly in the VGT expression by a yellow background, allowing the user to inspect how the instance is matched to its template. This is a very useful feature when inspecting instances that correspond to two or more ambiguous gestures, as the user can now clearly see the cause of the ambiguity. Current pitch values of numbered pitch frames are shown below the expression.

4 User Study

The aim of the user study was to find out whether the designers could understand VGT expressions, and to demonstrate the usefulness of the tool described in the previous section. Eight designers were recruited to participate in the study. Each participant (mean age=29.6, SD=2.8) had some previous experience with NVVI – four of them knew the interaction method, three had used it at least once and one had previously designed an NVVI application. Seven of the participants considered themselves as interaction designers and the remaining one as a usability expert. All participant were familiar with regular expressions.

The participants were given approximately 20 minutes of training, which involved discussing the syntax of two VGT expression examples as well as semantic errors. The participants were asked to complete three tasks. In each task they were told to recognize the gesture templates in given VGT expression by describing them orally and sketching a graphic representation of each template. They were also asked to identify any semantic errors that may have been present in the expressions and to propose a solution for each. However, they were not told to write a new corrected expression due to limited time of each session. One session lasted approximately one hour. Participants were divided into two groups of four – Group A and B. *Group A* was allowed to use the tool described above, whereas *Group B* was not allowed to use any aid.

Task #1

In the first task participants were told to analyze the following VGT expression:

```
p1 p* (p2 [p2.m - p1.m > 4] p* s <alpha> |
       p3 [p2.m - p2.m > 8] p* s <bravo>)
```

The expression above describes the two templates as depicted in Figure 5a. The *alpha* template defines instances where pitch increases by 4 or more midi notes. The *bravo*'s instances have to increase by 8 midi notes. However, the *bravo* template is unreachable, as the condition in the *alpha* template is always matched earlier.

Group A (Use of tool)**:** Each participant correctly understood the templates and discovered that the gesture *bravo* was unreachable. Two participants proposed a partially correct solution.

Group B: One participant misunderstood the *bravo* template and consequently could not see an error. The other participants miscategorized the error as ambiguous. Two participants proposed a partially correct solution.

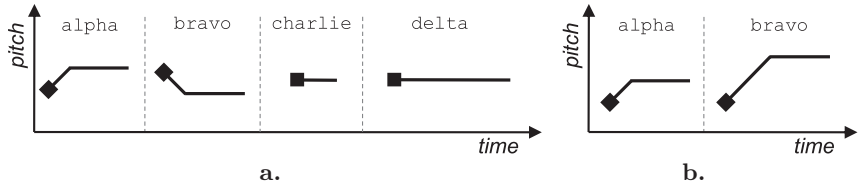


Fig. 5. a. Gesture templates in the task #1 **b.** Gesture templates in the task #2

Task #2

The second task contained gestures used in the *Tetris* game controlled by humming [7]. The participants were again told to analyze the VGT expression:

```
p1 p*
   (p2 [p2.m - p1.m > 4] p <alpha>* s |
    p3 [p1.m - p3.m > 4] p <bravo>* s) |
p*200;600 s <charlie> |
p*500; s <delta>
```

The expression above describes the templates depicted in Figure 5b. *Alpha* instances have to increase in pitch by 4 or more midi notes, whereas the *bravo* instances have to decrease by the same amount. *Charlie* instances are short tones of 200 to 600 ms and *delta* instances are all those that are longer than 500 ms. Two ambiguities are present in the expression. The first one is a time overlap in *charlie* and *delta* templates. The solution is to modify one of the limits. The second error is a pitch overlap between *alpha*, *bravo* and *charlie*, *delta* templates, due to the latter two not defining a pitch limit. The solution is to limit the pitch in *charlie*, *delta* templates to within ±4 midi notes.

Group A (Use of tool)**:** Each participant understood the presented templates. One participant incorrectly identified the gestures initially, but corrected their interpretation after using the tool. All four were also able to locate all errors and propose a correct solution for each error.

Group B: Unlike the three others, one participant was not able to describe *alpha* and *bravo* templates correctly. All four participants were able to find ambiguity between *charlie* and *delta*. The second error was found by three participants, who proposed a correct solutions for each of the errors.

Task #3

The most complex VGT expression was analyzed in the last task. The expression defines six of the eight templates used in keyboard controlled by humming [6].

```
p11 [p11.m< 60] p12 [p12.m-p11.m<=4 & p11.m-p12.m<=4]* s<alpha> |
p21 [p21.m>=60] p22 [p22.m-p21.m<=4 & p21.m-p22.m<=4]* s<bravo> |
p31 [p31.m< 60] p* p32 [p32.m-p31.m>4] p* s<charlie> |
p41 [p41.m>=60] p* p42 [p41.m-p42.m>4] p* s<delta> |
p51 [p51.m< 60] p* p52 [p52.m-p51.m>4] p* p53 [p53.m<=p51.m] p* s<echo> |
p61 [p61.m>=60] p* p62 [p61.m-p62.m>4] p* p63 [p63.m>=p61.m] p* s<foxtrot>
```

The six instances correspond to the following - 1. *alpha* to a straight low tone, 2. *bravo* to a straight high tone, 3. *charlie* to increasing tone by more than 4 midi notes, 4. *delta* to decreasing tone by more than 4 midi notes, 5. *echo* to a tone that increases by more than 4 midi notes and then decreases to at least its initial pitch and finally 6. *foxtrot* which is essentially *echo* vertically inverted. Ambiguities between *charlie* and *echo* and between *delta* and *foxtrot* are present due to the end pitch of *charlie* and *delta* templates not being limited.

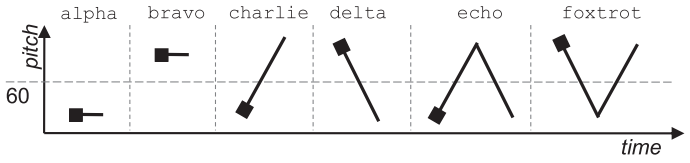


Fig. 6. Gesture templates in the task #3

Group A (Use of tool)**:** One participant misunderstood *alpha* and *bravo* templates. Two participants incorrectly identified the templates initially, but corrected their interpretations after using the tool. Two participants thought there was an error present between *alpha* and *bravo*, but identified their mistake after using the tool. All participants located the error and three of them were able to propose correct removal solution.

Group B: Two participants incorrectly identified *alpha* and *bravo* templates as unreachable and were thus unable to sketch them. The other two participant incorrectly identified the templates as ambiguous. However, the other templates were understood by all participants, who were also able to identify the ambiguities and propose correct solutions.

5 Discussion

Using VGT expressions accelerates the process of building an NVVI application, as the matching algorithm no longer needs to be hard coded. The question, that is raised though, is whether designers are able to understand these VGT expressions. In most cases, participants from both groups correctly identified templates directly from VGT expression, which supported our assumption that VGT expressions can be understood by most designers. From total of 48 gestures that were examined in one group, there was two errors in the group A (use of tool) and seven error in the group B. What was slightly surprising was that participants from group A primarily relied on their own judgement rather than on the provided tool. However, they did use the tool from time to time to visually confirm their opinion or when they were unsure of the answer. In these situations the tool helped them to correctly understand the given templates and consequently to succeed in fulfilling the tasks. Thanks to the tool, participants from the group A also had no difficulty in detecting semantic errors. Although

the participants from the group B were not as successful as group A, they were still able to locate a significant number of error occurrences. It seems that the use of the tool results in better understanding of VGT expressions and minimizes the overlooking of semantic errors. However, a further quantitative study is needed in order to support this hypothesis.

6 Conclusion

This paper discusses the formal description of pitch-based vocal input, used during the design process of NVVI applications. We have created a tool for automatic error detection and visualization of the formal description. Our research was focused on the comprehension of the formal description by designers and their ability to detect possible semantic errors with and without using the tool. Their ability to comprehend the formal description and to detect semantic errors was validated in a user study by eight interaction designers. Designers who used the tool were more successful in understanding the formal description. Further research concerning these results will be conducted in the future, including a comparative quantitative study to prove the efficiency of the gesture visualization tool.

Acknowledgments. This research has been partially supported by the MSMT research program MSM 6840770014 and the VitalMind project (IST-215387).

References

1. Harada, S., Landay, J.A., Malkin, J., Li, X., Bilmes, J.A.: The vocal joystick: evaluation of voice-based cursor control techniques. In: Proceedings of ASSETS 2006, pp. 197–204. ACM Press, New York (2006)
2. Hämäläinen, P., Mäki-Patola, T., Pulkki, V., Airas, M.: Musical computer games played by singing. In: 7th International Conference on Digital Audio Effects, pp. 367–371 (2004)
3. Igarashi, T., Hughes, J.: Voice as sound: using non-verbal voice input for interactive control. In: Proceedings of UIST 2001, pp. 155–156. ACM Press, New York (2001)
4. MIDI Manufacturers Association: Complete MIDI 1.0 Detailed Specification v96.1, 2nd edn. (2001), `http://www.midi.org/techspecs/midispec.php`
5. Poláček, O., Míkovec, Z., Sporka, A.J., Slavík, P.: New way of vocal interface design: Formal description of non-verbal vocal gestures. In: Proceedings of the CWUAAT 2010, pp. 137–144. Cambridge Press, UK (2010)
6. Sporka, A.J., Kurniawan, M., Slavík, P.: Non-speech Operated Emulation of Keyboard. In: Designing Accessible Technology, pp. 145–154. Springer, Heidelberg (2006)
7. Sporka, A.J., Kurniawan, S.H., Mahmud, M., Slavík, P.: Non-speech Input vs Speech Recognition: Real-time Control of Computer Games. In: Proceedings of ASSETS 2006, pp. 213–220. ACM Press, New York (2006)
8. Sporka, A.J., Žikovský, P., Slavík, P.: Explicative Document Reading Controlled by Non-speech Audio Gestures. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 695–702. Springer, Heidelberg (2006)

Appendix G

Humsher: A Predictive Keyboard Operated by Humming

Polacek O., Mikovec Z., Sporka A., Slavik P.: Humsher: a predictive keyboard operated by humming. : In The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility (ASSETS'11). New York: ACM, 2011, p. 75-82. ISBN 978-1-4503-0920-2.

Humsher: A Predictive Keyboard Operated by Humming

Ondřej Poláček Zdeněk Míkovec Adam J. Sporka Pavel Slavík
Faculty of Electrical Engineering, Czech Technical University in Prague,
Karlovo nám. 13, 12135 Praha 2, Czech Republic
{polacond, xmikovec, sporkaa, slavik}@fel.cvut.cz

ABSTRACT
This paper presents Humsher – a novel text entry method operated by the non-verbal vocal input, specifically the sound of humming. The method utilizes an adaptive language model for text prediction. Four different user interfaces are presented and compared. Three of them use dynamic layout in which n-grams of characters are presented to the user to choose from according to their probability in given context. The last interface utilizes static layout, in which the characters are displayed alphabetically and a modified binary search algorithm is used for an efficient selection of a character. All interfaces were compared and evaluated in a user study involving 17 able-bodied subjects. Case studies with four disabled people were also performed in order to validate the potential of the method for motor-impaired users. The average speed of the fastest interface was 14 characters per minute, while the fastest user reached 30 characters per minute. Disabled participants were able to type at 14 – 22 characters per minute after seven sessions.

Categories and Subject Descriptors
H.5.2 Information interfaces and presentation: User Interfaces – Input devices and strategies; Keyboard.

General Terms
Measurement, Performance, Design, Experimentation, Human Factors.

Keywords
Non-verbal Vocal Interface, Assistive Technology, Text Input, Predictive Keyboard, Adaptive Language Model

1. INTRODUCTION
Research in the field of text entry methods has been widely documented for some time. In static desktop environments we can observe the dominance of QWERTY keyboard which is caused by its extreme popularity rather than its optimal performance. Learning a new layout is a tedious process that can take more than 100 hours [1]. However, in special circumstances (e.g., impaired users, mobile environment) no dominant text entry method can be identified. This has consequently led to the development of many non-traditional approaches, where users accept longer learning time.

The maximum realistic text entry speed can be defined as a speed of an experienced typist using ten fingers on QWERTY keyboard. The speed will be approximately 250-400 characters per minute (CPM) for a professional typist [2]. With this speed achieved there is a little space for any enhancements like predictive completion, dynamic layouts, etc. as this will effectively slow down the type rate.

Physically disabled people usually cannot achieve such high speed due to their constraints. Their communication with computers is rather limited to only several distinctive stimuli – small number of physical buttons, joystick, eye-tracking, features of the electroencephalographic (EEG) signal etc. This limitation can be compared to a situation when we are typing with one finger only on virtual keyboard displayed on a touch screen. There is a research available [3], showing that typing with one finger on a touch screen with virtual QWERTY keyboard results in a speed 160 CPM for expert users after 30 minutes training. If we reduce the size of the virtual keyboard to 7 cm then the speed will drop to 105 CPM. The speed reached by physically disabled people will be certainly lower. This situation opens a space for research of new entry methods which will take into account various limitations of motor impaired users and increase the entry speed.

There is currently a range of assistive tools available to help users with motor impairments. However, each user may have significantly different capabilities and preferences according to the range and degree of their impairment. In case of severe physical impairment, people usually have to use other interaction methods to emulate the keyboard. One of the methods that has been successfully used by people with special needs is the non-verbal vocal interaction (NVVI) [4]. It can be described as an interaction modality, in which sounds other than speech are produced, for example humming [27] or vowels [28].

Our virtual keyboard, *Humsher*, described in this paper utilizes vocal gestures, i.e. short melodic and/or rhythmic patterns. The user can operate the keyboard by humming. Each key is assigned a pattern. It has been designed for those people with upper-limb motor impairments such as quadriplegia induced from stroke, cerebral palsy, brain injury etc. Additionally, users are required to have healthy vocal folds enough to be able to produce humming. The main advantages of such interaction are its language independence and fast and accurate recognition as opposed to speech [4]. Speech recognition software usually works relatively well for native speakers; however, the accuracy is much lower for accented speakers or for people with speech impairment.

ASSETS'11, October 24–26, 2011, Dundee, Scotland, UK.

1.1 Definitions of Terms

Probably the most common measures of performance of text entry methods are *words per minute* (WPM) or *characters per minute* (CPM) [29]. Both rates indicate speed of a text entry method. Relation between them is defined by Equation 1. ISO 9241-4 standardizes WPM rate for keyboards at CPM divided by five, i.e. one "word" is considered as five characters including spaces. CPM is defined by equation 2, where |T| is length of written text in characters and S is time in seconds.

$$WPM = \frac{1}{5} \times CPM. \quad \textbf{(1)}$$

$$CPM = \frac{|T| - 1}{S} \times 60. \quad \textbf{(2)}$$

A *gestures per character* (GPC) rate [29] is also used in this paper for evaluating purposes. Gesture is regarded as an atomic operation. In the case of the humming input, vocal gestures are treated as atomic operations. Text entry methods with low GPC rate are considered as better than those with high rate; however, other parameters must be taken into account, such as length or complexity of the vocal gesture. The GPC rate is defined by Equation 3, where $|IS_{\square}|$ is an input stream which contains all vocal gestures produced by the user and |T| is length of written text in characters.

$$GPC = \frac{|IS_{\emptyset}|}{|T|}. \quad \textbf{(3)}$$

A sequence of *n* characters is referred to as *n-gram*. The n-grams with length equal to one, two and three character are being called unigrams, bigrams and trigrams respectively. In the paper the term *n*-gram is used for strings of characters of an unspecified length *n*.

2. RELATED WORK

There is a wide range of text entry methods targeting the motor-impaired users. We can notice that the methods described in this section often differ significantly in physical interaction used, which is determined by specific motor impairment. Each method is often unique for concrete impairment conditions and thus it typically makes no sense to compare various methods as they are not in concurrent position. Several principles can be identified in the literature – *predictive completion*, *ambiguous keyboards* and *scanning*.

A text entry method can be accelerated by prediction, when a list of possible completions is updated with each entered character. This reduces number of keystrokes per character. The Reactive Keyboard [5] predicted possible words according to context that had been already written. An adaptive dictionary-based language model was used. Predicted candidates could be selected by the mouse cursor. Expert users of a QWERTY keyboard would be slowed down, however, such prediction is useful for poor typist or people with limited movement of upper limbs. Another predictive keyboard GazeTalk [6] predicted six most probable letters and six words according to current context. If no prediction was correct, there was full keyboard available. This virtual keyboard was controlled by the eye gaze. The keys were activated by dwell-time selection system [7]. The average typing rate achieved by novice users was 16 CPM.

Probably the most prevalent ambiguous keyboard is the commercial T9 system by Tegic Communications [8] that is widely adopted on mobile phones. The idea behind is simple – the alphabet is divided into nine groups of characters and then each group is assigned to one key. The user selects desired characters by selecting the keys and after a sequence of keys is entered the word is disambiguated using a dictionary. For its efficiency, similar ambiguous keyboards were designed for physically impaired people. Kushler [9] describes an ambiguous keyboard in which the alphabet was assigned to seven keys and the eighth key was used as a space key that initiated the disambiguation process. Tanaka-Ishii [10] published similar system, in which only four physical keys were used. Besides disambiguation, the text entry method was capable of predicting words. The average speed of this method was 70 CPM, achieved after ten sessions by able-bodied participants. Harbusch [11] presented similar method in which the whole alphabet was assigned to only three keys and one key was used for executing special command in a menu.

When the number of stimuli, which can be issued by the user, is limited to only one or two, using scanning technique is inevitable. For example in the case of two buttons, the first button can be pressed repetitively (scanning) to select a key and second button is used to confirm the selection. When only one button is available, the keys are selected automatically for a certain amount of time. After the time expires, next key is selected. The button is used to confirm the selection again. Keys can be spatially organized in a matrix and the desired key is then selected by row-column scanning [12]. Combining linear scanning with an ambiguous keyboard is a common technique. For example, Kühn [13] used four-key scanning ambiguous keyboard and achieved 35 CPM without out-of-vocabulary words. Miro [14] limited the number of keys to only two (keys 'a-m' and 'n-z') and estimated its entry rate to 50 CPM for an expert user. Beltar [15] used three keys and developed a virtual mobile keyboard. In QANTI [16] three keys are mapped to the alphabet. The keyboard is operated by one switch that is triggered by intentional muscle contractions. The typing rate ranges from 12.5 to 33 CPM.

An efficient system is Dasher [17], which is based on a dynamically modified display and adaptive language model [18]. The characters are selected by moving the mouse cursor around the screen. Continuous "one finger" gestures are used as the input method. This is a very suitable input method for motor impaired users, who can operate a pointing device. The writing speed achieved is approximately 100 CPM with experienced users reaching up to 170 CPM. For users who have no hand function, a modification of the Dasher system can be made to allow input via eye tracking. A longitudinal study [19] found that an average writing speed of 87 CPM after ten 15-minutes sessions could be achieved. This speed was a large increase from the initial speed of just 12.5 CPM. Speech Dasher [20] is another interesting modification of Dasher. It combines speech input with the zooming input of Dasher. The system must first recognize a user's utterance. Errors are then corrected via the zooming input. Expert users reached a writing speed of approximately 200 CPM.

Sporka et al. [21] describe the NVVI-based method of keyboard emulation. Each vocal gesture is assigned a specific key on the keyboard, when a gesture is produced a corresponding key is emulated. The average reported typing rates varied between 12 and 16 CPM, which was measured in a study with able-bodied participants. Different assignments of NVVI gestures to keys were investigated, namely the pitch-to-address, pattern-to-key and Morse code mappings. In the pitch-to-address mapping, the

keyboard was mapped onto a 4×4×4 matrix, while a sequence of three tones of specific pitches determined an address in the matrix. In the pattern-to-key mapping each key was assigned a specific gesture.

Another keyboard operated by NVVI is CHANTI [25]. It is an ambiguous keyboard, where the alphabet is split into only three groups. The keyboard combines philosophy of the ambiguous keyboard QANTI [16] and humming input. Scanning technique is replaced by direct selection of a key by vocal gestures. The keyboard was tested with five severely motor-impaired people, the speeds ranged from 10 to 15 CPM after 7 sessions.

Additionally, the P300 speller [22] is a method that utilizes the electroencephalographic (EEG) signal in the human brain to control a virtual keyboard. The keyboard is a 6x6 matrix containing alphanumeric characters. The user focuses on a character and as the character flashes, the brain produces a stimulus. At least two flashes are needed to input a character. According to Wang et al. [22], the writing speed achieved is approximately 7.5 CPM.

3. HUMSHER DESIGN

Our virtual keyboard, *Humsher*, has been designed for severely motor-impaired people, who can control it by vocal gestures. It utilizes the same language model as Dasher [17] (prediction by partial match; PPM [18]). The model provides n-grams and their probability, which have been predetermined by a given context. The model is initialized from a small corpus of English text, but it adapts as the user types.

3.1 Dynamic Layouts

The interfaces described in this section employ dynamic layout. The n-grams, which are extracted from the PPM model, are offered sorted according to their probability. The probability is predetermined by already written text. Practically it means that after typing an n-gram, the context is updated, probabilities of following n-grams are recounted and the layout is displayed accordingly.

We designed and implemented three different user interfaces (Direct, Matrix and List) with dynamic layout of characters. Each interface differs in either vocal gesture set or in mapping of gestures to actions. The Direct and Matrix interfaces utilize six vocal gestures as depicted in Fig. 1, whilst the List interface utilizes only three simple vocal gestures as depicted in Fig. 2. The vocal gestures are explicitly identified by its length (short/long) or by its pitch (low/high). In order to distinguish low and high tones a threshold pitch needs to be adjusted for each user – e.g. the difference between male and female voice is as much as one or two octaves. Only two different pitches were chosen as with increasing number of pitches, more precise intonation is required and the interaction becomes more error prone.

All three interfaces offer n-grams, containing the characters how the text might continue, sorted according to the probability. The n-grams can be unigrams (individual characters) as well as bigrams, trigrams, etc. The length of n-grams is not limited, only probability matters. N-grams to display are chosen according to the following steps:

1. Add all unigrams to the list L that will be displayed.
2. For each n-gram in the list L compute probability of all (n+1)-grams and add them to the list L if their probability is higher than a threshold.
3. Repeat step 2 until no n-gram can be added.
4. Sort the list L according to probability of each n-gram.

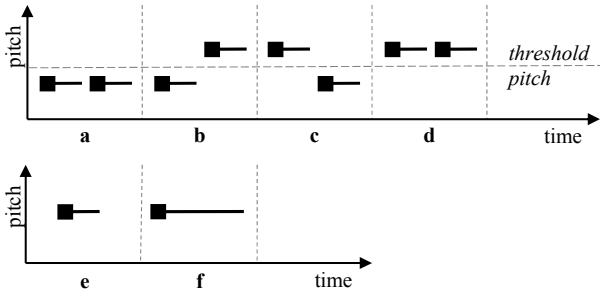


Figure 1. Vocal gestures used in Direct and Matrix interfaces

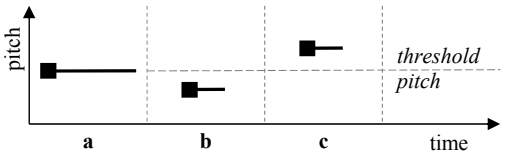


Figure 2. Vocal gestures used in List and Binary interfaces

3.1.1 Direct interface

The *Direct* interface (see Fig. 3) allows users to directly choose from four cells (labeled cell 1 to 4) in the Active column (part A). These cells contain n-grams that have been determined as the most probable following characters of the written text. Cells can be selected by vocal gestures depicted in Fig. 1a-d:

a. two consequent low tones (cell 1),
b. a low tone followed by a high tone (cell 2),
c. a high tone followed by a low tone (cell 3),
d. two consequent high tones (cell 4).

If there is no cell in the Active column that contains the desired character, the user has to move the leftmost column in the Look ahead (part B) to the Active column by producing a single short tone (see Fig. 1e) and keep repeating it until the desired n-gram appears in one of the cells in Active column. Text, which has been already written, can be erased by producing a long tone (see Fig. 1f). The longer the user keeps producing the tone the faster are the characters erased.

3.1.2 Matrix interface

The *Matrix* interface (see Fig. 4) utilizes the same vocal gestures as the Direct interface, however, the user interaction is different. Users are presented with a 4×4 matrix of the most probable n-grams. Cells in the left column of the matrix contain the highest probable n-grams, whilst the rightmost cells contain the lowest probable n-grams.

Selection of the correct cell is accomplished in two steps by specifying a column and a row. First, the user must select a column by producing a corresponding vocal gesture (Fig. 1a-d). The column is then highlighted and the same vocal gestures can be used to select the desired cell by selecting a row. If a character does not appear in the matrix, the user has to produce a short tone (see Fig. 1e) in order to display less probable n-grams. Written text can be erased by producing a long tone (see Fig. 1f), in the same manner as in the Direct interface.

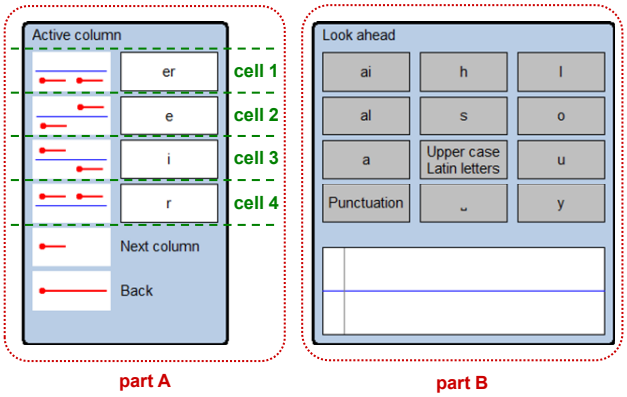


Figure 3. Direct interface. A– active column, C B – look ahead matrix

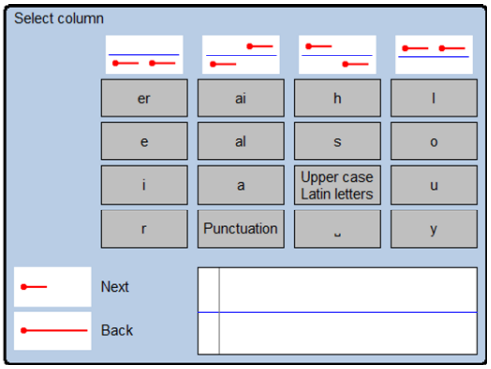


Figure 4. Matrix interface

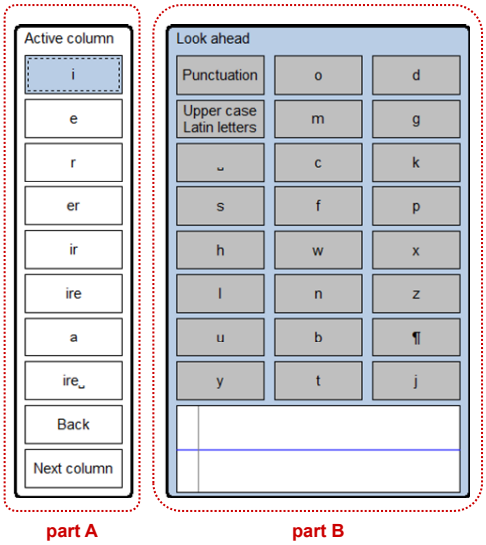


Figure 5. List interface. A – active column, B – look ahead matrix

3.1.3 List interface

The *List* interface (see Fig. 5) is controlled by just three simple and easy-to-learn gestures (see Fig. 2). The Active column (part A) presents the user a list of cells containing the eight most probable n-grams. The topmost cell is selected. Users can move the selection up and down by producing a short high or low tone (see Fig. 2b,c). A long tone (see Fig. 2a) is used to confirm the desired selection. This interface does not utilize special vocal gestures to select the next column or erase written text. Instead, these two functions are always made available by introducing two special cells Back and Next column at the bottom of the Active column list.

3.2 Static Layout

Static layout was designed in order to simplify the process of visual location of desired character. In dynamic layouts users have to locate a character visually by linear scanning and they cannot rely on the visual memory. The process of locating correct character can be tedious for low-probable characters. Moreover, users sometimes do not notice a correct character and they have to rotate through the whole list of characters and n-grams once again. This consequently can lead to users' frustration. Therefore we decided to implement a static interface, called Binary interface, that keeps position of characters and the characters are sorted alphabetically. Time needed to locate a character is then modeled by Hick-Hyman law [24] and it is logarithmically dependent on the length of the alphabet. Locating characters visually in the static layout is obviously faster than the same task in dynamic layouts as logarithmic scanning is used instead of linear.

3.2.1 Binary interface

In the *Binary* interface (see Fig. 6) the characters are always displayed in an alphabetic order. Such order gives us an opportunity to select desired character by binary search algorithm adopted from basic programming techniques. The algorithm locates position of a character in the alphabet by splitting it into two halves and deciding which half is used in the next step. Then the half is split again and again until the correct character is found. Each character is located in following number of steps:

$$steps = \lceil \log_2 N \rceil \qquad \textbf{(4)}$$

N is size of the alphabet. In our case the algorithm would require $\lceil \log_2 36 \rceil = 6$ selections as our alphabet contains 36 symbols. The user would have to produce six vocal gestures to enter a character. Therefore the best theoretical GPC rate achieved by the binary search is equal to six, which is quite high. But what happens if the alphabet is split according to the probability of characters rather than into two exact halves? Then a character with high probability could be located in fewer steps, however, character with low probability might be located in even more than six steps. The actual GPC rate measured empirically in a user study presented later is much lower than six.

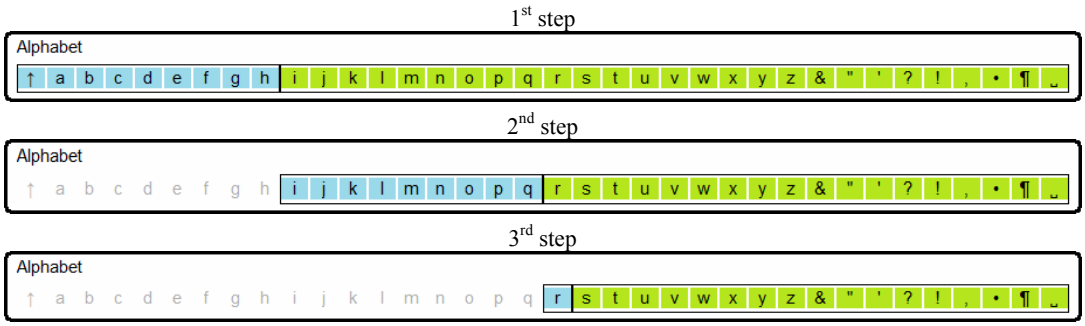


Figure 6. Binary interface, typing "*r*" after "*Text ent*"

The Binary interface is based on modified binary search algorithm. In each step the alphabet is split into two groups with balanced probability, i.e. the sum of probabilities of characters in each group is as close to 0.5 as possible. The boundary between groups is then computed according to the Equation 5, where k is the index of boundary character, p_i is a probability of character i and N is a size of the alphabet.

$$\min_{1<k<N} \left(\left(\frac{1}{2} - \sum_{i=1}^{k-1} p_i \right)^2 + \left(\frac{1}{2} - \sum_{i=k}^{N} p_i \right)^2 \right). \qquad (5)$$

The Binary interface utilizes only three vocal gestures (see Fig. 2) as well as the List. Short low tone (Fig. 2b) and short high tone (Fig. 2c) are used for entering text, while the long tone (Fig 2a) is used for corrections.

An example of user interaction with the Binary interface is depicted in Fig. 6. Let us assume that the user has already entered the text "Text ent" and wants to continue by entering character "r". In the first step the alphabet is split into two groups "shift -h" and "i-space". The user chooses the second group by producing a high short tone. In the second step the rest of the alphabet is split into groups "i-q" and "r-space". Again the second group is chosen by the same high short tone. In the last step "r" is the only character in the first group because of its high probability. Remaining characters are in the second group. The character "r" is now entered by low short tone. In this case the character was selected only in three steps by three short tones.

When comparing Binary interface to the other three interfaces, several features can be observed:

- User can easily locate desired character as letters are sorted alphabetically and characters do not change their positions while entering text.
- Simple vocal gestures are employed (similar to List interface). Only two gestures are used for entering text and one for deleting text.
- The Binary interface offers only single characters unlike the interfaces with dynamic layout. It is not possible to enter more characters at once.

4. EVALUATION

In order to evaluate the interfaces we conducted two user studies. The goal of the first one was to compare all four interfaces, measure their speed and find out user's opinions on them. In the second study four disabled participants were recruited to validate potential of Humsher for motor-impaired users.

4.1 Comparison of interfaces

The aim of the user study was to measure the writing speed of each interface and subsequently determine which interface was the most efficient. In the study 17 able-bodied participants (10 men, 7 women, mean age=26, SD=2.1) took part. Each participant completed four sessions. According to Mahmud et al. [23], four sessions are needed to minimize the error rate of the NVVI. The schedules of each session are outlined below:

- **Session 1:** Participants were trained in producing the required vocal gestures. After reaching an accuracy of 90%, they were presented with all interfaces and asked to enter short phrases with each of them. This session lasted approximately 30-60 minutes depending on the user's abilities.
- **Sessions 2 and 3:** Participants were asked to enter two simple phrases using all interfaces. The sessions were conducted remotely and they lasted roughly 20 minutes.
- **Session 4:** Participants were asked to enter three phrases using all interfaces. The session was conducted remotely and it lasted roughly 30 minutes. Objective data from this session were collected.

After the last session each participant performed a subjective evaluation of each interface by means of remote interview. The participants received approximately 24 hours rest between the sessions. In order to minimize the learning effect, the sequence of interfaces was counterbalanced. Objective results (CPM, GPC rate and number of corrections) are shown in Table 1.

Table 1. Means and standard deviations (SD) of the typing rate (CPM), vocal gesture per character (GPC) rate and total number of corrections.

Interface	CPM		GPC		Corrections	
	Mean	SD	Mean	SD	Mean	SD
Direct	14.4	2.8	1.8	0.23	13.0	11.0
Matrix	11.8	2.1	1.9	0.32	16.1	14.6
List	13.0	3.2	3.5	0.58	6.4	6.6
Binary	11.7	1.8	3.4	0.18	14.5	8.5

The ANOVA test and Scheffé's method [26] were used to find statistically significant differences in mean quantities among interfaces. When comparing mean CPM rates, the Direct interface was significantly faster ($F(3,67) = 4.20$, $p < .01$) than the Matrix interface and it was also significantly faster than the Binary interface. Other differences in speed were not significant.

In the case of List and Binary interfaces, the users had to produce significantly more ($F(3,67)= 107.7$, $p < .01$) vocal gestures per

character than Direct and Matrix interfaces. This corresponds to number of vocal gestures used in the interfaces. Direct and Matrix interfaces utilize six complex gestures (see Fig. 1), while the other interfaces only three simple gestures (see Fig. 2). As mentioned in section 3.2.1, theoretical GPC rate for standard binary search is 6, when the alphabet contains 36 symbols. By modifying the binary search, we succeeded to reduce the GPC rate to 3.4 empirically measured in the user study.

After the last session, participants were asked to comment on the interfaces. The Direct interface was mostly perceived as accurate and fast. The Matrix interface was in many cases perceived as fastest among all interfaces, although it was slower than Direct and List interfaces. Additionally, the List interface, which is not the slowest, was reported as the slowest. The List interface was also reported as cumbersome – some participants complained that it was not transparent enough and the navigation was tedious. This is probably due to the high number of cells in columns, which makes the visual searching more difficult. The Binary interface was found easy and fast by most participants, although it was the slowest one. The participants appreciated static layout of the interface, however, eight participants complained about the fact that only one character can be entered at one time and the method does not offer n-grams as the dynamic layout interfaces. The participants also made positive comments on simplicity of vocal gestures used to control the interface. Although there were no significant differences in objective data between List and Binary interfaces, participants strongly preferred the Binary one.

We identified two main searching strategies employed by participants when using Direct and List interfaces. Some of them visually scanned only the first column (Active column, see Fig. 3 and 4). When searched character was not found in this column, they moved forward and scanned the first column again. Some of them also reported that the Look ahead matrix is redundant and confusing. The other participants visually scanned all cells in Active column and Look ahead matrix. When searched character was not found, they moved forward and scanned the last column. They reported that this strategy allows them to plan vocal gestures in advance, which they found faster.

Ten participants reported fatigue of vocal folds during the experiment, which they mostly compensated for by lowering their pitch and dropping their voice.

Table 2. Performance of expert users

Interface	Expert 1			Expert 2			Expert 3		
	CPM	GPC	corr	CPM	GPC	corr	CPM	GPC	corr
Direct	29	1.5	1	24	1.7	8	30	1.5	2
Matrix	23	1.6	3	20	1.9	15	23	1.5	2
List	25	2.8	0	17	3.4	4	26	2.9	1
Binary	23	3.6	1	16	3.6	23	20	3.2	10

4.1.1 Typing rate of expert users

Learning a new text entry method is always a long-term process. The study presented results of novice users, who were given only necessary amount of training. In order to determine possible upper limit of performance of all Humsher interfaces, three experienced NVVI users were given 4-6 hours of training. The typing rate was recorded after their performance did not improve significantly. Table 2 summarizes CPM, GPC rates and number of corrections for each interface. The speed varied between 16 and 30 CPM. Expert 1 and 3 preferred the Direct, while expert 2 preferred Matrix interface.

4.2 Case studies with disabled people

The goal of the study was to find out whether Humsher can serve as an assistive tool for motor-impaired people. Four people were recruited in cooperation with local non-profit associations. The study was longitudinal, it was organized in seven sessions and each session lasted 30-60 minutes. First, the participants were asked to use the Binary interface because of its simple vocal gestures. Then they were asked to learn more complicated gestures and use the Direct interface, because it was the fastest one. The rough schedules of each session are outlined below:

- **Session 1:** The participants were asked to describe how they use ICT and how they enter text. Then they were trained in producing vocal gestures starting with the easiest ones (see Fig. 2). Binary interface was presented and the participants were asked to enter a phrase.
- **Session 2:** Participants trained more complicated vocal gestures (see Fig.1) until required accuracy was achieved. Then the Direct interface was presented to them and they were asked to enter a few phrases.
- **Session 3 – 7:** Participants were asked to enter phrases using the Direct interface. On the last day the participant were asked to describe experience using the interfaces.

While training the vocal gestures, the thresholds for low/high and short/long tones were personalized for each user. Two users with speech impairments were not able to consciously alter pitch of their tone, therefore a new gestures were designed especially for them.

4.2.1 Participant 1

The participant was 30 year old IT specialist in a small company, quadriplegic since birth. Due to privacy protection, he only participated in the study remotely. We conducted interviews with him via telephone and e-mail.

He uses a mouth stick to operate his PC (keyboard and mouse). Apart from the Sticky Keys tool available in Microsoft Windows he uses no other assistive technology. He uses various system administration tools, word processors, graphic and sound editors and he feels no disadvantage in comparison with other users.

He found the Direct interface precise and pleasant to use. Overall, he said he felt in control when using the tool. “The system allowed me to write whatever I wanted. I was not forced into any options.” He used the word “intelligent” to describe the suggested options provided by the tool when typing text. He achieved a mean type rate of 22 CPM. He reported, however, that his current text entry rate achieved by the mouth stick is higher.

4.2.2 Participant 2

Another disabled participant was 19 years old, quadriplegic since an accident about 3 years ago. He is a high-school student who uses computer to access study materials, talk with his friends over text media (especially e-mails), make telephone calls and watch movies. He spends typically 2 to 4 hours using his laptop equipped with NaturalPoint SmartNav4 head motion tracker and Click-N-Type keyboard emulation software. However, he is able to use the head motion tracking system only for 2-4 hours and then he gets too tired. He had a previous experience with another NVVI based interface for entering text.

When working with Binary interface, his mean type rate was 12 CPM. After switching to Direct interface, the type rate increased to 21 CPM. Although he was almost two times faster with the

Direct interface, he reported that the Binary interface was quicker and more responsive ("I like that it is fast. I can see it all in front of me and I know exactly what to do next."). He felt more in control than when using the Direct interface ("I am a bit lost when using the Direct interface as I sometimes do not notice the right option."). The participant considered our method similar in speed to his current assistive technology and he would use it as an alternative solution when his head gets too tired.

4.2.3 Participant 3

The participant was a 58 year old woman with cerebral palsy. All her limbs are affected by the disease. She can sit on a chair, but she needs a wheelchair for movement. She has a lot of unintentional movements in her arms. Her voice is also affected. She speaks slowly and she does not articulate properly. Her health state is slowly but steadily declining.

She used to work as an office staff in a non-profit organization, but she is unemployed for one year now. She used to type on a typewriter and a computer keyboard. However, now her performance decreases and she types very slowly on a keyboard. The only assistive technology that she uses is a trackball to control the mouse pointer. She also tried speech recognition, but it did not work for her at all.

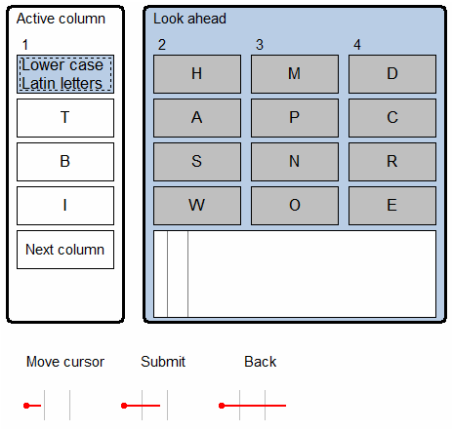


Figure 7. Modified List interface

She spent first and second sessions trying to learn voice gestures for the Binary interface. However, after two sessions she could hardly write a phrase. She was not able to effectively alter pitch of her tone, which led to many corrections. Therefore the vocal gestures were changed to short, medium long and long tone. Then she was asked to use it for another two sessions and she reached 8 CPM.

As the participant was unable to produce more complicated gestures, we modified the List interface (see Fig. 7) for use with the new gesture set. Short tone was used to move cursor in the Active column down, medium tone to submit selected n-gram and long tone for correction. She used this interface for remaining three sessions and reached 15 CPM.

The participant reported that the speed of the modified List interface is similar to her current typing rate and she was interested in purchasing it as a product. She also made comments on speech recognition ("This is much better than speech for me"). She reported that after one hour of humming her vocal chords were not tired at all.

4.2.4 Participant 4

The participant was 51 years old, quadriplegic since an accident about 22 years ago. His legs and right arm are paralyzed. He can use his left arm to operate wheelchair, however, fine motoric of his left hand is reduced. His vocal chords and neck muscles are also slightly affected.

Before the accident he used to work as a machine engineer. Since that he is unemployed. He has never worked with computers, but he regularly uses cell phone for couple of years, mainly for calling and writing short text messages. However, composing message is a tedious process for him.

The participant started with Binary interface and used it for two sessions. He experienced similar problems to participant 3. As he was not able to produce low and high tone properly, his performance was about 1 CPM with a lot of corrections. In the third session he switched to the modified List interface (see Fig. 7) as participant 3 and his performance increased rapidly with minimum mistakes. Using this interface and the vocal gestures based on length he reached type rate of 14 CPM.

He stated that typing text with Humsher is faster and better than typing on his cell phone. Generally he was pleased with the modified List interface. However, his vocal chords got tired after 40 minutes of humming.

5. CONCLUSION

This paper has presented and evaluated four interfaces of Humsher – an adaptive virtual keyboard operated by humming. Three of them (Direct, Matrix and List) used dynamic layout, in which characters were sorted according to its probability. The layout was updated after entering a character. The last interface (Binary) used a static layout, in which characters were displayed alphabetically and did not change their position. A character was selected by modified binary search algorithm that took into account probability of each character.

Most novice users preferred the Binary interface, even though it was not the fastest one. They appreciated mostly the static layout of characters and simple vocal gestures used to control the interface. On the other hand expert users preferred interfaces with dynamic layouts. Interfaces with dynamic layout were perceived worse, however, users appreciated that sometimes several characters could be entered together. The Direct interface was the fastest one with average speed 14.4 CPM achieved by novice and 28 CPM by expert users.

Acceptance of our tool for the target group was verified by the inclusion of four motor-impaired participants. Two of them could not use speech recognition software as their speech was also impaired. Cases of all disabled participants are described separately in a longitudinal and qualitative study. Their speed achieved after seven sessions varied between 14 and 22 CPM.

While some techniques, such as Dasher [19], offer their users type rates up to 100 CPM, they may not be used by people with severe motor impairments without expensive hardware, such as eye trackers. Our method requires no additional hardware to a standard PC and performs better than the NVVI Keyboard [21] and CHANTI [25] methods which have the identical hardware requirements and for which a similar performance is reported: 16 CPM for NVVI Keyboard, 15 CPM for CHANTI, and 22 CPM for Humsher.

6. ACKNOWLEDGMENTS

This research has been supported by the MSMT research program MSM 6840770014 and the Veritas project (IST-247765).

7. REFERENCES

[1] Silfverberg, M. 2007. Historical Overview of Consumer Text Entry Technologies. In Text Entry Systems: Mobility, Accessibility, Universality, I. S. MacKenzie and K.Tanaka-Ishii (eds.). Morgan Kaufmann, 3-26.

[2] West, L. J. 1998. The Standard and Dvorak Keyboards Revisited: Direct Measures of Speed. http://samoa.santafe.edu/media/workingpapers/98-05-041.pdf, Technical report, Santa Fe Institute.

[3] Sears, A., Revis D., Swatski, J., Crittenden, R., Shneiderman, B. 1993. Investigating touchscreen typing: the effect of keyboard size on typing speed. In J. Behaviour and Information Technology, vol. 12, 17-22.

[4] Igarashi, T., Hughes, J F. 2001. Voice as sound: using non-verbal voice input for interactive control. In Proceedings of UIST ’01, ACM Press, 155-156.

[5] Darragh, J. J., Witten, I. H., James, M. L. 1990. The Reactive Keyboard: A Predictive Typing Aid. Computer Journal, vol. 23, IEEE Press, 41-49.

[6] Hansen, J., Johansen, A., Hansen, D., Itoh, K., Mashino, S. 2003. Language technology in a predictive, restricted on-screen keyboard with ambiguous layout for severely disabled people. In Proceedings of EACL Workshop on Language Modeling for Text Entry Methods.

[7] Jacob, R. J. K. 1990. What you look at is what you get: eye movement-based interaction techniques. In Proceedings of CHI’90, ACM Press, 11-18.

[8] Grover, D. L., King, M. T., Kushler, C. A. 1998. Reduced keyboard disambiguating computer, Technical report, US Patent Publication.

[9] Kushler, C. 1998. AAC: Using a Reduced Keyboard.

[10] Tanaka-Ishii, K., Inutsuka, Y., Takeichi, M. 2002. Entering text with a four-button device. In Proceedings of the 19th International Conference on Computational Linguistics, Association for Computational Linguistics, 1-7.

[11] Harbusch, K., Kühn, M. 2003. Towards an adaptive communication aid with text input from ambiguous keyboards. In: Proceedings of EACL’03, Association for Computational Linguistics, 207-210.

[12] Simpson, R., Koester, H. 1999. Adaptive one-switch row-column scanning. In IEEE Transactions on Rehabilitation Engineering, vol. 7, no. 4, IEEE Press, 464-473.

[13] Kühn, M., Garbe, J. 2001. Predictive and highly ambiguous typing for a severely speech and motion impaired user. In Proceedings of 1st International Universal Access in Human-Computer Interaction Conference, UAHCI ’01.

[14] Miro, J., Bernabeu, P. 2008. Text entry system based on a minimal scan matrix for severely physically handicapped people. In Computers Helping People with Special Needs, LNCS 5105, Springer, Heidelberg, 1216-1219.

[15] Belatar, M., Poirier, F. 2008. Text entry for mobile devices and users with severe motor impairments: handiglyph, a primitive shapes based onscreen keyboard. In Proceedings of ASSETS ’08, ACM Press, 209-216.

[16] Felzer, T., MacKenzie, I., Beckerle, P., Rinderknecht, S. 2010. Qanti: A software tool for quick ambiguous non-standard text input. In Computers Helping People with Special Needs, LNCS 6180, Springer, 128-135.

[17] Ward, D. J., Blackwell, A. F., MacKay, D. J. C. 2000. Dasher – a data entry interface using continuous gestures and language models. In Proc. of UIST ’00, ACM, 129-137.

[18] Teahan, W. 1995. Probability estimation for PPM. In: Proceedings of the New Zealand Computer Science Research Students’ Conference.

[19] Tuisku, O., Majaranta, P., Isokoski, P., Räihä, K. J. 2008. Now Dasher! dash away!: Longitudinal study of fast text entry by eye gaze. In Proceedings of the 2008 symposium on Eye tracking research and applications, ETRA ’08, ACM Press, 19-26.

[20] Vertanen, K., MacKay, D. J. 2010. Speech dasher: fast writing using speech and gaze. In Proceedings of CHI ’10, ACM Press, 595-598.

[21] Sporka, A. J., Kurniawan, S. H., Slavík, P. 2006. Non-speech operated emulation of keyboard. In: Clarkson, J., Langdon, P., and Robinson, P. (eds.) Designing Accessible Technology, Springer, London, 145-154.

[22] Wang, C., Guan, C., Zhang, H. 2005. P300 brain-computer interface design for communication and control applications. In: Proceedings of the 27th Annual International Conference of the Engineering in Medicine and Biology Society, IEEE-EMBS’05, 5400-5403.

[23] Mahmud, M., Sporka, A. J., Kurniawan, S. H., Slavik, P. 2007. A Comparative Longitudinal Study of Non-verbal Mouse Pointer, In Proceedings of INTERACT 2007, Springer, Heidelberg, 489-502.

[24] Hyman, R. 1953. Stimulus Information as a Determinant of Reaction Time. Journal of Experimental Psychology, vol. 45, 188-196.

[25] Sporka, A. J., Felzer, T., Kurniawan, S.H., Polacek, O., Haiduk, P., MacKenzie, I.S. 2011. CHANTI: Predictive Text Entry Using Non-verbal Vocal Input. In proceedings of CHI’11, ACM Press., 2463-2472.

[26] Maxwell, S.E., Delaney, H.D. 2004. Designing Experiments and Analyzing Data: A Model Comparison, ISBN 0805837183, Lawrence Erlbaum Associates, 217–218.

[27] Sporka, A.J., Kurniawan, S. H., Slavik, P. 2004. Whistling user interface (u3i) In 8th ERCIM International Workshop “User Interfaces For All”, LCNS 3196, Springer, 472-478.

[28] Harada, S., Landay, J.A., Malkin, J., Li, X. and Bilmes, J.A. 2006. The Vocal Joystick: Evaluation of voice-based cursor control techniques. Proceedings of ASSETS '06,ACM Press, 197-204.

[29] Wobbrock, J.O. 2007. Measures of text entry performance. In Text Entry Systems: Mobility, Accessibility, Universality, I. S. MacKenzie and K.Tanaka-Ishii (eds.). Morgan Kaufmann, 47-74.