ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2 Issue 1 Ver 2 Jan-March 2014
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

An Eyes-Free Model Implementation: Voice Based Optical Character Recognition for Mobile Devices

[I]Preethi.P, [II]Baskaran.G, [III]Christo Paul.E
[I,III]2nd Year-M.E CSE, Srinivasan Engineering College, Perambalur, Tamil Nadu, India.
[II]Assistant Professor, Dept. of IT, Srinivasan Engineering College, Perambalur, Tamil Nadu, India.
Email : [I]preethi1.infotech@gmail.com, [II]gbaskarancse@gmail.com, [III]echristopaul@gmail.com

Abstract
Research into eyes-free mobile reading devices has grown in recent years. This research has focused mainly on the image processing required by such a device, with a lower emphasis on the user interaction. In this paper, a model of a voice user interface (VUI) for a mobile reading device is presented. Three field studies with blind participants were conducted to develop and refine the model. Particularly, blind people can use these devices resorting to screen reading software commonly along with a painless exploration approach. However, for a large part of the population, a first approach to these devices is still challenging and a future reigned by touch surfaces presents as daunting. The success with such gadgets varies and is highly dependent on the user's abilities. Small displays on mobile handheld devices, such as PDAs and cellular phones, are the bottlenecks for usability of most content browsing applications. Generally, conventional content such as documents and web pages need to be modified for effective presentation on mobile devices. A formal grammar is used to describe the VUI, and a stochastic Petri net was developed to model the complete user-device interaction. Evaluation and analysis of the user testing of a prototype led to empirically derived probabilities of grammar token usage for the commands that comprise the VUI .

Keywords
Audio user interfaces, human computer interaction, human factors, interactive systems, modeling, speech recognition, stochastic systems, user interfaces.

I. Introduction

Human interaction with automated control systems that employ modes. We focus our discussion on the features of the control system that lead to mode confusion and error. We first provide working definitions of the term "mode" and discuss key constructs that contribute to mode error, such as mode ambiguity and user's expectations. Following this, we discuss human interaction with automated control systems, in general, and cockpit automation, in particular. To provide a formal description of human interaction with such control systems, we introduce a modelling language, which is used by system engineers to specify complex control systems. We use the language to describe the three types of modes that are commonly found in modern control systems: interface, functional, and supervisory modes. In modern control systems, a mode is a common architecture for grouping several machine configurations under one label. The set of modes in a control system corresponds to a set of unique machine behaviours. The operator interacts with the machine by switching among modes manually, or monitoring the automatic switching triggered by the machine.

In this paper our focus has been on the features of the machine that may lead to mode confusion and error in user-machine interactions. On the machine side we have discussed the constructs of mode behaviour and mode ambiguity. On the user side we have discussed user expectations, which are influenced by three factors: the user's task, the user's knowledge of the machine's behaviour, and the user's ability to sense triggering events. But no systematic evaluation of human interaction with automated control systems can take place unless some representation is available. The building blocks of the representation suggested in this paper are the language and the three modelling structures (interface, functional, and supervisory). The representation is oriented toward the user's perspective (but it can be refined to include more detail). Now we can combine our understanding of the constructs that lead to mode error with a formal description of control systems. The first question that can be asked is whether the display allows the user to discern between machine configurations that are part of the user's task. Second, one can compare the user's model of the machine's behaviour with these configurations, and ask whether the user's model makes it possible to predict the next configuration of the machine. Last, one can ask whether the display provides the user with all the information necessary to reliably predict when a transition will take place. Such analysis can highlight mismatches between the behaviour of the machine and the corresponding information provided to the user via the interface (e.g., display) and training materials (e.g., manuals). We believe that the mode constructs described in this paper and a State Machine modelling language such as State charts, provide the foundation and basic tools for systematic and formal evaluations of human interaction with control systems. The continuing search, we argue, must be for a systematic and formal analysis that will allow designers to identify the potential for mode error earl y in the design phase.

In most complex control systems include modes, and their use will only increase in the foreseeable future. Mode confusion and error contribute to mishaps and fatalities in the operation of these systems. Only a combination of methodologies, some from very different disciplines, can provide us with effective resources to address this problem. This is our challenge.

The goal of the work presented in this article is the development of a stochastic Petri net (SPN), for use in the application development of a VUI that supports a mobile reading device. To understand the important qualities of a reading device for the blind, one must understand the reading task, and to note that the goal of reading is comprehension. Many reading models have been developed; a simple model is comprised of four fundamental perceptual scanning, text processing, word recognition, and comprehension. The perceptual scanning step refers to the way that the human eye senses the characters of a written page. Text processing involves character recognition, which leads into word recognition. By maintaining the context of what is read in short term memory,

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

readers are able to match the recognized words with the context and improve the overall comprehension of the presented text.
By considering how a sighted person reads a newspaper, three design implications can be derived for the development of a mobile reading device.

Regression

While this term refers to a reader's ability to easily reread a portion of text (which occurs 10 to 15% of the time for sighted readers), a device should provide the ability to easily navigate throughout the text – forward or backward in large and small segments.

Spatial Cues

Sighted readers utilize many spatial and temporal cues to find specific sections within a document [18]. The spatial dimension, however, is removed from a user of an auditory device. Thus, a device must supplement this temporal information with a form of spatial information.

Find-ability

Without visual cues, a device must provide other forms of navigation, including audible cues for a user to find information within a document.
The model developed here begins by considering the human subject as the complete system with input from a long cane and audible sounds. The subject's goal is to safely navigate to predetermined destination. By starting with this initial model, an eyes-free navigation model for reading is proposed in which a reading device, and the human subject are considered as the entire system. This human-in-the-loop model lays the foundation for further refinement, which is modelled as a Rasmussen decision ladder and a hierarchical task analysis of the interaction.

A hierarchical task analysis is presented. In this analysis, the system oriented tasks are on the left, and the user-oriented tasks are on the right. The system performs the document image processing tasks that convert the document image into text. This text is then processed by the text-to-speech engine, which will in turn read the text to the user. The system uses text-to-speech processing to produce system commands that guide the user in the navigation of the document. The system also performs speech-to-text processing to evaluate and respond to user commands. The user portion of the task analysis is broken into three parts: listening, evaluation, and navigation. The listening task encompasses the hearing of text or commands from the system. The evaluation tasks involve understanding the text and commands to determine whether the human subject should continue listening, respond to a command, or issue a new command. Finally, the navigation task is comprised of the user issuing voice commands to the system, or a response to commands from the system.

II. Related Work

Research in the area of adaptation of documents to different output devices and allocation of document information to different output channels has diverged into at least two directions. One direction concerns how to reformat document content for small devices, transforming information represented in one visual information channel into another visual channel. The other direction concerns making document information accessible to visually impaired users by transforming visual information into audio information. In the visual-to-visual transformation category, some solutions focus mainly on readability of text, displaying some text in larger fonts and allowing users to select page elements to be zoomed in or collapsed, summarizing the text content, semantic grouping and re-flowing content based on reading order. Solutions that focus both on text and images include Enhanced Thumbnails and Smart Nails. Enhanced Thumbnails contain keywords extracted from the source document and pasted onto a low-contrast down sampled page image. Smart Nail technology creates an alternative image visualization for a single document page by scaling, cropping, and reflowing page elements, subject to display size constraints. Both techniques include image and text, but the output is a static visual representation of each page.
In Multimedia Thumbnails, the output consists of document information from multiple pages represented in a dynamic way using animation and audio. Some of the most relevant prior art to our current work in the visual-to-visual category is described in, where a method for non-interactive picture browsing on mobile devices was proposed. The goal there was to find salient, face and text regions on a picture automatically and then apply zoom and pan motions on this picture to automatically provide informative close ups to the user. This method concentrates on representing photos, whereas our method focuses on representing high-resolution multi-page document content. Moreover, the automated picture browsing technique shows only visual information, whereas we employ visual and audio channel for communicating document information.
Designing an efficient hierarchical auditory menu system has long been recognized as a difficult problem. These challenges are revealed in the design of automatic interactive voice response (IVR) systems, which are often referred to as "touchtone hell" Incremental improvements to IVR menus have been proposed to ease user frustration. The earPod technique is designed for an auditory device controlled by a circular touchpad whose output is experienced via a headset. When a user touches the dial, the audio menu responds by saying the name of the menu item located under the finger. Users may continue to press their finger on the touch surface, or initiate an exploratory gesture on the dial. Whenever the finger enters a new sector on the dial, playback of the previous menu item is aborted. Boundary crossing is reinforced by a click sound, after which the new menu item is played. earPod is designed to allow fast expert usage. The weaknesses of traditional voice menu designs are first analyzed. The limitations of the audio modality compared to the visual one are well understood. The visual modality allows displaying a list of choices instantaneously and persistently. Touchpads arguably have a richer input vocabulary than keypads because they allow gliding gestures in addition to discrete taps.
When the information is uninteresting, users can skip to an adjacent area. Or they can listen to the entire message and repeat it if needed. It is especially useful to be able to promptly switch to a new item before the previous one finishes playing because users often understand partial audio messages. Overall, the audio technique had an accuracy of 92.1%, while the visual technique yielded 93.9% accuracy. Although the earPod technique can be used in many situations, it is particularly suited to mobile use. The circular touchpad could be embedded into devices of many different form factors or it could be implemented as a separate component, like a wireless remote control. Touchpads are typically very light, allowing a touchpad remote to be carried easily on a neck lanyard or a key chain. While some custom design will undoubtedly be needed to adapt earPod to such individual devices,

the effort required to execute the designs should be relatively straightforward.

Blind and visually-impaired people cannot access essential information in the form of written text in our environment (e.g., on restaurant menus, street signs, door labels, product names and instructions, expiration dates). In this paper, we present and evaluate a mobile text recognition system capable of extracting written information from a wide variety of sources and communicating it on-demand to the user. Architecture. a visually-impaired business-woman who values her independence. Before going to work in the morning, Jane needs to take her allergy medications. Not remembering when she had renewed her prescription, she reaches for her camera phone and takes a picture of the bottle to check the expiration date. The phone reads out the expiration date, which happens to be last month. Jane decides to throw the bottle away and to make a detour to the pharmacy. At work, Jane schedules a lunch meeting with a client. After she walks to the restaurant, she snaps a picture of the street sign at the corner of the block to verify that she is at the correct intersection. As she enters the restaurant, Jane realizes that a small object is blocking her path. To address this problem, users can leverage familiar conventions to find common locations for text (e.g. placards by doors and street signs at intersections). Users can also point the camera-phone directly at objects they hold in their hands (e.g., restaurant menus and product packaging). The vast majority of the visually-impaired can rely on their existing, albeit low, vision capabilities to locate text. Alternatively, our prototype can be used in conjunction with other systems that help blind people detect objects around them. snap a photo of an object and automatically send it to a server using an HTTP request over the GPRS network; a server-side script invokes the OCR engine, which extracts the text from the image. The server sends the extracted text back to the phone, where it is displayed and enunciated using a speech-synthesis engine. Our approach shares a similar architecture, but we use an off-the-shelf camera-phone instead of a PDA to better address the needs of visually-impaired users. Instead of requiring the user to specify where text is in the image (which is nearly impossible for our users),we simplify the interaction by removing this extra step. Several other research systems and commercial products use onboard cameras and the processing power of PDAs for text extraction. Our approach requires the user to carry only an ordinary camera-phone and demonstrates that special-purpose hardware is unnecessary. Some of the challenges described in this paper may disappear in the future as the smart phones will be equipped with better lenses and cameras and as good text-recognition software becomes available for the phone platforms. The proof of concept outlined in this paper shows that this important problem can be addressed effectively using existing technologies

III. Our System And Assumptions

System components and relations.

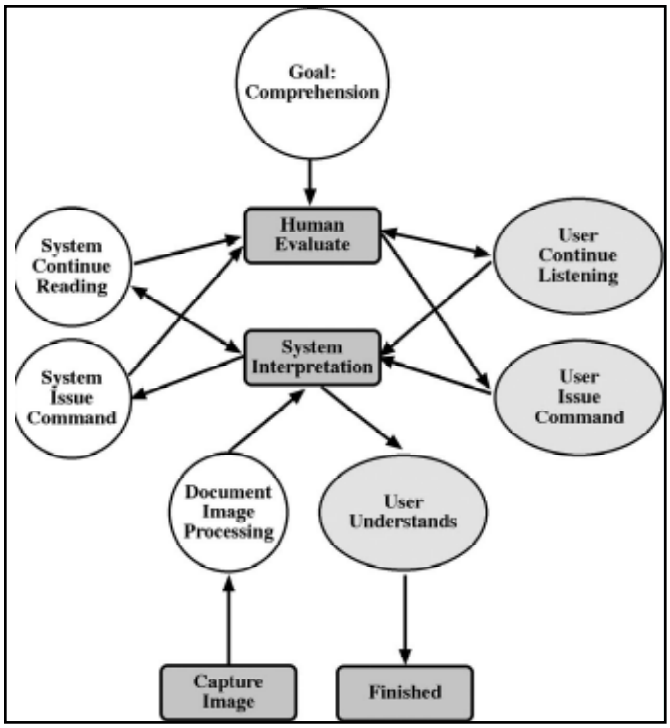

Fig. 1: Architecture.

While this human-in-the-loop model provides a strategic understanding of the system, it glosses over the important details of the interaction and is only able to be evaluated at a conceptual level. Thus, this model is refined with a Rasmussen decision ladder, as shown in Figure. Briefly, the decision ladder is a generic representation of the steps that may be involved in decision making. In this model, the system and the human are modeled together. The goal of the user is represented at the top as comprehension. The system-oriented tasks are presented as circles on the left-hand side of the model, and the user-oriented tasks are presented as ovals on the right. The system will process the document image, read the document to the user, and issue directional commands to the user when needed. The user will listen to the document or issue commands to the system based on the evaluation of what is heard, moving toward the goal of comprehension.

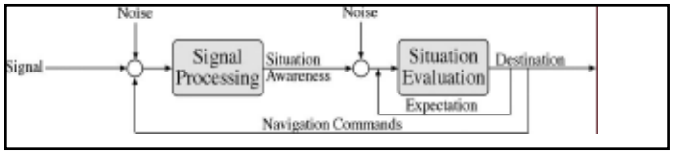

Fig. 2: Generic Eye Model.

Image capture

A document images contain typically a large amount of high frequency data such as text. text in documents is meant to be read in a predetermined reading order and image and text units. In this way, both recognizability and navigation problems are addressed by having text readable and figures comprehensible after zooming and panning and minimizing the navigational input required by the user.

Data transfer

The image document to be transfer to recover .The content matching is a description of the document content that contains semantic labels of document elements, such as title, section heading, and

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

caption, as well as their absolute coordinates on the page.

Receiver(text to speech)

Text-to-speech (TTS) technology on a computer refers to the combination of text appearing on the computer display together with the computer speaking that text aloud with a digitized or synthesized voice. Digitized speech is a recorded (or digitized) human voice speaking, and synthesized voice is a computer-generated voice speaking the text. This tutorial focuses on TTS software tools that use synthesized speech to read any text.

Sender(speech to text)

Recognizing the speaker can simplify the task of translating speech in systems that have been trained on specific person's voices or it can be used to authenticate or verify the identity of a speaker as part of a security process. Speech-to-Text reproducing speech into a text format onto a computer screen at verbatim speeds .

Language identification

Automatic Language Identification (Language ID) is the problem of identifying the language being spoken from a sample of speech by an unknown speaker. The human is by far the best language ID system in operation today.

IV. System Preliminaries

When the desired outputs are available, the text-to-speech module synthesizes the information to the user. The speech-to-text processing for interpretation of user commands.

A. Optical character recognition

Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic conversion of scanned images of handwritten, typewritten or printed text into machine-encoded text. It is widely used as a form of data entry from some sort of original paper data source, whether documents, sales receipts, mail, or any number of printed records. It is a common method of digitizing printed texts so that they can be electronically searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

Modern general-purpose speech recognition systems are based on Hidden Markov Models. These are statistical models that output a sequence of symbols or quantities. HMMs are used in speech recognition because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal. In a short time-scales (e.g., 10 milliseconds), speech can be approximated as a stationary process. Speech can be thought of as a Markov model for many stochastic purposes. Speech recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition" Transcription of speech into text, video captioning.

B. A novel Dynamic time warping

Dynamic time warping is an approach that was historically used for speech recognition but has now largely been displaced by the more successful HMM-based approach. Dynamic time warping is an algorithm for measuring similarity between two sequences that may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video the person was walking slowly and if in another he or she were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics – indeed, any data that can be turned into a linear representation can be analysed with DTW.

Since the large majority of Smartphone devices do not have a physical keyboard, typing is enabled by the on-screen QWERTY keyboard that appears on the device screen. This operation is time-consuming and error-prone for visually impaired users. Two main families of text entry techniques have been proposed for the visually impaired. One family is based on the idea to allow the user to navigate through the characters using, for example, directional gestures. The second family of techniques is based on Braille. The process of entering a key with Voice-over is divided into two stages: identification of the target key and its confirmation. In the identification stage, the blind user scans the keyboard searching for the target key while the names of the touched keys are read by speech as soon as the keys are touched. Five main problems are identified with experiments conducted with blind users.

(1) Since a single key has a small size, also a skilled blind user can rarely identify a target key without first exploring the keys around it.
(2) In the target key confirmation stage it is relatively frequent that the user trying to enter a key inadvertently slides the finger over another close-by key, thus causing a typing mistake.
(3) Moving among layers is time consuming as it additionally requires to identify and confirm the key to change the layer.
(4) No shortcut exists to enter the keys that are used more frequently like the \blank space" and the delete".
(5) The input technique is fully dependent on speech output. TypeInBraille enables the user to input a character through its Braille representation by inserting the three rows of each cell from the top to the bottom.

A tap on the left part of the screen corresponds to the left dot raised and the right dot flat. Typing efficiency in terms of words per minute is not a meaningful metric by itself to evaluate typing performance. One observation about the accuracy is that, in particular in the second task, many users committed mistakes due to the fact that they typed blank spaces when not needed.

The feedback from the users helps us understand why the performances of TypeInBraille are less degraded in an uncomfortable environment with respect to the on-screen QWERTY. There are two main reasons. First, typing with the on-screen QWERTY strongly relies on the speech feedback that is only partially audible in the \tramcar" scenario.

V. Experimental Evaluation

The experimental results were analyzed using analysis of variance (ANOVA), with repeated measures on both the interaction and the question type, and an α value of 0.05. ANOVA results showed compared to the sequential reading, the navigable reading significantly reduced response time ($F(1, 7) = 5.59$, $p = 0.05$) and improved satisfaction ($F(1, 7) = 94.74$, $p < 0.0001$).

The interaction effect of the interaction mode and question on the response time is also significant ($F(1, 7) = 20.36$, $p = 0.003$). The main effect of question on the response time is marginally significant ($F(1, 7) = 3.29$, $p = 0.11$). Summarize the mean of participants' response time and satisfaction for each of four treatment combinations, respectively.

An unexpected discovery was that the mean response time to find

the answer to the question that was not related to the headline was smaller for the sequential reading than for the navigable reading. This may be due to the fact that the question for the navigable reading contained the word "Obama" as did the headline for article three. Thus, three of the eight participants requested article three before requesting article one. While this is not conclusive, it does indicate that for some tasks a sequential reading can be faster than a navigable reading. Further research would be needed to confirm this. Participants were asked to rate the difficulty of answering the posed questions using a four-point Likert scale questionnaire.

Each participant's opinion was ranked with 1 being the best and 4 being the worst. Though the sequential reading may have been faster for some of the participants, all of the participants rated the navigable interaction higher than the sequential reading During the post-test survey, each participant was asked his or her opinion of the completeness of the proposed grammar. Half of the participants suggested adding the term "back" to simply go back one sentence or paragraph as the proposed "repeat" command does. Similarly, they suggested supplementing the "skip" command with a default to advance to the next sentence rather than require the user to say the word "sentence."

Recall that two important qualities of a mobile reading device for the blind are the provision of spatial cues and find-ability. To support the spatial cues and find-ability, the system must track the location of one document image relative to the others in the same document and therefore direct the user with in document navigation. This feature is supported with audible system commands. In this section, the grammar for the system commands is presented, followed by a formal evaluation of the commands and results.

It is interesting to note that all six participants assumed that the first line of text was at the top of the page and felt for the top of the page to take the first picture. However, the text on many sample pages started further down the page, and thus the camera needed to be moved closer to the participant. During the post-test survey, participants indicated that the audible prompts were useful in completing the task and provided an improved interaction.

One interesting note is that three of the six participants wanted to know how far to move the device when prompted with a movement direction. The ability to provide this feature, however, is beyond the scope of this research. The system command grammar withstood the test of the users and therefore did not need to be updated.

VI. Conclusions

The development of a VUI for a blind user in the use of mobile reading devices that supports regression, find-ability, and provides spatial cues was described in this paper. The model development began with a generic model of eyes-free navigation, which was then modified to accommodate eyes-free navigation for reading. Automatic Language Identification (Language ID) is the problem of identifying the language being spoken from a sample of speech by an unknown speaker. This SPN was evaluated and refined further, culminating in a proposed model that could be used to guide the implementation of a VUI for mobile reading devices.

VII. Acknowledgement

First and foremost, The authors would like to thank the God Almighty, who guides us always in the path of knowledge and wisdom.We thank the editors and anonymous reviewers for their valuable comments to significantly improve the quality of this paper.We are very much grateful to all the staff members and my friends who helped a lot to complete this work.

References

[1] S. Brewster, "Using non-speech sounds to provide navigation cues," ACM Trans. Computer Human Interact., vol. 5, no. 2, pp. 224–259, 1998.

[2] F. Brooks, Mythical Man-Month: Essays on Software Engineering. Reading, MA: Addison-Wesley, 1995.

[3] P. A. Carpenter and M. Dehneman, "Lexical retrieval and error recovery in reading: A model based on eye fixations," J. Verbal Learn. Verbal Behav., vol. 20, no. 2, pp. 137–164, Apr. 1981.

[4] T. Dumitras, M. Lee, P. Quinones, A. Smailagic, D. Siewiorek, and P. Narasimhan, "Eye of the beholder: Phone-based text-recognition for the visually impaired," in Proc. Int. Symp. Wearable Comput., 2008, pp. 145– 146.

[5] G. Ghiani, B. Leporini, and Ú. F. Patern, "Supporting orientation for blind people using museum guides," in Proc. CHI Extended, 2008, pp. 3417–3422., pp. E3-1–E3-10.

[6] Robert Keefer, Yan Liu, and Nikolaos Bourbakis, "The Development and Evaluation of an Eyes-Free Interaction Model for Mobile Reading Devices" Vol. 43, No. 1 January 2013.

Preethi.P received the B.Tech. degree in Information Technology from Roever Engineering College affiliated to Anna University, Tamil Nadu, India, 2012 . She joined Srinivasan Engineering College affiliated to Anna University in 2012 for the M.E. degree in Computer Science and Engineering. Her research interests include affective computing, mobile computing, networking and applied cryptography.

Baskaran.G received the B.E. and M.E. degrees in Computer Science and Engineering from Bannari Amman Institute of Technology, Tamil Nadu, India and University Department, Anna University, Coimbatore, Tamil Nadu in 2008 and 2011, respectively. He is an Assistant Professor in department of Information Technology and supervisor of M.E. students at Srinivasan Engineering College affiliated to Anna University. His research interests include affective computing, networking, cryptography, network security and cloud computing.

Christo Paul.E received the B.Tech. degree in Information Technology from Roever Engineering College affiliated to Anna University, Tamil Nadu, India, 2012 and also he got 37th university Rank in that academic year of 2012 . He joined Srinivasan Engineering College affiliated to Anna University in 2012 for the M.E. degree in Computer Science and Engineering. His research interests include affective computing, cloud computing, mobile computing and applied cryptography.