Author(s): Lahti, Lauri & Kurhila, Jaakko

Title: Low-cost portable text recognition and speech synthesis with generic software, I

Year: 2007

Version: Post print

Please cite the original version:

Lahti, Lauri & Kurhila, Jaakko. 2007. Low-cost portable text recognition and speech synthesis with generic software, I. Human Computer Interaction International 2007, 22-27 July 2007, Beijing, China. P. 918-927. ISSN 1611-3349 (electronic). ISSN 0302-9743 (printed). ISBN 978-3-540-73280-8 (printed). DOI: 10.1007/978-3-540-73281-5_100

Note: The final publication is available at link.springer.com

Lahti, L., & Kurhila, J. (2007). Low-cost portable text recognition and speech synthesis with generic software, laptop computer and digital camera. Proc. Human Computer Interaction International 2007, Vol. 6 (Universal Access in Human-Computer Interaction – Ambient Interaction), 22-27 July 2007, Beijing, China (ed. Stephanidis, C.), LNCS 4555, 918–927. Springer. ISBN 978-3-540-73280-8. http://link.springer.com/chapter/10.1007%2F978-3-540-73281-5_100

Low-Cost Portable Text Recognition and Speech Synthesis with Generic Laptop Computer, Digital Camera and Software

Lauri Lahti [1] and Jaakko Kurhila [2]

[1] Department of Computer Science and Engineering, P.O. Box 5400,
FIN-02015 Helsinki University of Technology, Finland
Lauri Lahti at oi fi
[2] Department of Computer Science, P.O. Box 68, FIN-00014 University of Helsinki, Finland
kurhila at cs helsinki fi

Abstract. Blind persons or people with reduced eyesight could benefit from a portable system that can interpret textual information in the surrounding environment and speak directly to the user. The need for such a system was surveyed with a questionnaire, and a prototype system was built using generic, inexpensive components readily available. The system architecture is component-based so that every module can be replaced with another generic module. Even though the system makes partly incorrect recognition of text in a versatile environment, the evaluation of the system with five actual users suggested that the system can provide genuine additional value in coping with everyday issues outdoors.

Keywords: Text recognition, speech synthesis, independent initiative.

1 Introduction

Coping with everyday life is an important issue for everyone [14]. As the use of technology has increased in everyday life, visually challenged or blind people have encountered new challenges and a need for adaptation in their routines. On the other hand, emergence of technical solutions has offered new possibilities to be an active and independent member of the society despite of the loss of sight. Research on various aspects of augmenting the eye sight with technical innovations is ongoing (see e.g. a face recognition system for social interactions [6], Braille interpretation for persons unable to read Braille [11], and way-finding with Braille output [15]).

It is evident that transforming visual textual information to speech can be of value since especially in urban areas direct and indirect textual information about the surrounding environment is largely available. Purpose-built systems for transferring text to speech in outdoor environment are being developed (see e.g. [4, 1]).

Since we live in an era of technology, many individuals have already a relatively lightweight laptop computer and a digital camera. These generic components can be combined into a low-cost portable text recognition and speech synthesis for outdoor use, if the components are bound together with appropriate software.

In this paper, we briefly describe the results of a survey that motivated the need for such a generic portable combination, describe the system and report the results of its use and performance in outdoor situations. The design principle behind the system is that the construction of the application should be component-based and use software that is easily available. The discussion in the end sketches the direction of porting the system into a digital mobile phone.

2 Survey of the Needs

In order to survey the demand for low-cost assistive technology for coping in everyday life, an email questionnaire was sent out to 450 members of the Finnish Federation of the Visually Impaired. A total of 29 persons replied to the questionnaire. Half of them had a complete loss of sight, and rest of them had a faint ability to perceive light or shapes. They represented fairly evenly age groups from twenties to sixties. Even though the questionnaire examined various aspects of assistive technology with 94 separate questions [8], the results reported in this article concentrate only on two specific issues: independent initiative and portable assistive technology for visually impaired users. The first issue of independent initiative was examined with two questions: “Do you try to cope with everyday problems by asking help from others or reading independently by yourself?” and “Would you like to manage your everyday activities more independently and how it could be the most beneficial for you?”

Several respondents state that they try to cope with the problems independently, but if they fail (after reasonable efforts), help from other people is sought. The justifications for this vary from “not wanting to be of trouble” to “lack of courage to seek help” and “not wanting outside people to know my personal affairs”.

A respondent concludes that ”[...] of course I would like to cope with my everyday life as independently as possible. It is fairly tedious to work out schedules in order to get a guide to run errands. In my opinion, I would be more equal with others if I could run my errands on my time, and not when a family member or an aid has time.”

The second issue of portable assistive technology was examined with a question: “Special needs are being met with pocket-sized computers to alleviate the problems of everyday life wherever the user goes. What kind of features would be beneficial for you in this kind of assistive device?”

Out of 26 replies, 13 respondents brought up the wish of speech usage. An excerpt of a reply describes the possibilities of assistive technology in this area: “The computer should have a small Braille display and possibly speech synthesis. One could use it, for example, with an ATM machine, in order to know what the screen says. Similarly, it could be used with other screens, e.g. at bus, subway and railway stations. The computer could help when coping with new routes and it could substitute as a map, if it told street names and directions to aim at with a guide dog after entering the final destination to the system.”

Another respondent summarizes general needs: “When moving around, it would be undoubtedly good. But at the same time, it should have all the other things as well, such as phone, notebooks, address books, the Internet [...] but it should be an existing device, so that every assistive feature is just an add-on. This way, the accessibility and

the price could be manageable. Nowadays the pricing of purpose-built assistive technology is out of reach. In addition, there are too many devices that provide only one or two services. Everything should be packaged into one portable device!"

After these results, it was clear that there is a need for a portable, low-cost solution to help in independent initiative that can serve multiple purposes. The idea of portable device for supplementing low vision or loss of sight is not particularly new; there already are various solutions [6, 7], and ongoing projects are under way [10]. Independent initiative in other contexts has also been researched [14].

The novel idea behind our system is that the construction of the assistive application should be based on devices and software that are already easily available — preferably freely downloadable — on the consumer market. The approach seemed to be cost-effective and provided an opportunity to tailor the assistive application with a large variety of modules. Without a doubt, existing software components combined in a novel way provides a considerable potential for a variety of computational tasks.

3 System Description

As machine vision is still limited in object recognition in everyday life [12, 2], the system was built to support only textual information, even though there are plenty of issues in textual recognition as well (see e.g. [3, 20, 19]).

3.1 Operation from the User's Viewpoint

After certain preparations the operation of the system is simple. The user points the camera to a view that needs to be interpreted and presses the left mouse button. The view is then captured by the camera and saved on the computer's hard drive. After that the image file is analyzed by a character recognition program. The text that can be found is transmitted to a speech synthesis program and the result can be heard from headphones. This procedure can be achieved with only one click with the mouse and the auditory interpretation of the texts in the scenery is acquired in 30 seconds.

The system searches one type of the characters at time: dark characters on light background or light characters on dark background. By rolling the wheel of the mouse forward the user can repeat the hearing of the current interpretation. If the user rolls the wheel of the mouse backwards the system offers interpretation made from the same picture but with inverted colors. By pressing the wheel of the mouse user can interrupt the hearing of the interpretation if it is necessary.

3.2 System Architecture

The final prototype of the portable system that provides text-to-speech synthesis in outdoor environment consists of mostly generic components: a laptop computer connected to a digital camera, easy-to-acquire software, a wheel-mouse and headphones. The laptop computer used was Toshiba Satellite Pro 4600 with a Pentium III processor (391 MHz). The camera was Canon PowerShot A95 with a CCD of 5 megapixels. The weight of the combination was less than 4 kilograms.

The operation of the system is based on the cooperation between software components running under Windows XP. The components used for the prototype

were: Remote Capture software by Canon, TopOCR character recognition software by Topsoft [17], Mikropuhe speech synthesis by Timehouse [16], and Winamp media player by Nullsoft [13]. Remote Capture makes it possible to capture images directly from a Canon digital camera to the computer. TopOCR offers means to perform character recognition on any JPG image file. Mikropuhe is one of the leading software for producing synthesized speech in Finnish.

The cooperation is conducted in Autohotkey [9] macro environment. Autohotkey offers a scripting language for describing the desired flow of actions and their conditions within the operating system. On the top of the Autohotkey environment, a script is needed to allow the user to control the flow of data between the camera, OCR and speech synthesizer software. The script needed for the purpose was designed and written by the first author. All the other software components are generic in a sense that they are not custom-built for assistive technology. Therefore, it should be noted that even though the components were not all open source or freely distributable software, comparable components can be acquired free of charge. The decision to use relatively expensive speech synthesizer software was a language-related issue. The component-based architecture allows using any useful or easy-to-acquire components.

4 Text Recognition with the System

The quality of interpretation of the texts in the surrounding environment varies significantly. Due to challenges in the character recognition process, the system can normally offer only a suggestive interpretation. Normally, the system captures excerpts of text and thus conveys only a selection of the original text to the user. In addition, it is typical that optical character recognition software interprets random visual elements as characters, so that the end result can be difficult to comprehend. Thus the visually impaired users should not rely solely on this information but instead use it as a supplement for other observations concerning environment. Despite the distortion, it is often possible to recognize familiar words even from very short excerpts. Awareness of the context and common sense reasoning still leads to understanding of the text-to-speech interpretation.

Example in Figure 1 shows the quality of the system output in interpreting textual input in a typical condition. Of course, interpretations transcribed on paper do not match the user experience when perceived with speech synthesis.

Figure 1 has been taken towards a fence at a construction site. On the fence there is a sign that says: *"Työmaa-alue. Asiattomilta pääsy kielletty."* (Construction site. Unauthorized access forbidden).

When this picture is analyzed by the system the text in this picture produces an interpretation: *".-.,X.3=Tyomaa-alueAsiattomiltapääsykiellettyåö."*

To eliminate confusing splitting of words the system concatenates all characters on purpose when producing the speech output. Especially in Finnish language, this should help in preserving the proper pronunciation and make the end result more understandable. Despite of the concatenation and some additional characters in the result, the original message is in practice quite recognizable. Even the loss of diacritic dots is tolerable to understand Finnish language. The design principle of using existing components forces to accept a certain level of robustness in the system.

Fig. 1. View at a construction site

5 User Tests

Two male and three female volunteers (aged 32 to 66) with varying visual disabilities tested the system in real life conditions. Two of the volunteers were not capable to read visual text at all. Three of the volunteers were able to read enlarged text with strong contrast. They all were relatively active users of computers.

During the testing the users were assisted by the first author. The assistance was for coordinating the activity and making a detailed recording of the opinions expressed. Some of the test locations were familiar to the volunteers.

5.1 Results

The evaluation consisted of a total of 35 different locations. In some locations, few additional trials were needed to optimize, for example, the framing of the textual information visible. In twenty locations the system could offer interpreted keywords that can be classified as “useful” or “rather useful”. In four other locations the interpretations can be classified as “slightly useful”. In the rest eleven locations, the system was not able to produce useful interpretations. When the experience of the system and its use grows, it is reasonable to believe that the ratio becomes better.

As it is often the case with speech synthesis, understanding the synthesized speech from the machine-made interpretation of the text was sometimes challenging. One source of inconvenience in this case is the intentional merging of the text into a long word, as motivated in Section 4. Another source for the difficulties appeared to be the special fonts used in many logos and advertisements. Moreover, random visual elements are often interpreted as characters. The defects in the interpreted texts were

considered annoying but, on the other hand, they are much the same in results obtained in all traditional scanning and optical recognition of visual text.

Biased interpretations of the text received varied reactions from the users of the system. Additional characters give sometimes a misleading impression of looking at a timetable or a price list. Numeral information gets easily an uncertain sequencing and the existence of dots and commas is unclear. The system produces classical confusions with recognizing letter "O" and number zero or small letter "L" and number one. The volunteers mention similar difficulties with letter "B" and number eight or Roman numbering. However, the volunteers were already used to cope with these uncertainties with common sense and contextual information. For example, familiar prefixes in telephone numbers and Web addresses help to recognize the correct type of information.

5.2 Examples of Test Cases

Figure 2 presents a view to an exit from a subway station. There is an exit sign that says: *"Raitiovaunut Spårvagnar"* (Trams, in Finnish and in Swedish).

The system produced the following interpretation of the text: *"teRaitiovaunutSparvagnar"*. The volunteer had no trouble in interpreting the output: *"Hey, that is the exit to go to tram tracks!"* The key issue is that the user can augment the output with existing knowledge to form a sense of the current context.

An opposing case to a successful interpretation in Figure 2 is presented in Figure 3. In Figure 3, there is a view towards a shop entrance: *"LAHJATALO PASTEL"* (GIFT HOUSE PASTEL, in Finnish).

The system produced the following interpretation of the text: *"LAHJATALOPASTEL"*. The volunteer's response for the speech output is: *"Can I have it again? [2nd listening] Lahjatalopastel? Maybe a missing letter? Lahjatalo..."*

It is apparent that the volunteer did not know the presence of this particular gift shop, and could not connect the name "Pastel" to previous knowledge from the environment to the context. In this case, the addition of a clear space between the "gift shop" and "Pastel" could have helped understanding.

Other user test scenarios and results are presented in [8], containing details about the test users' characteristics and the process of interpretation in varying outdoor environment.

5.3 General Comments of the Use and Development Ideas

The volunteers gave versatile feedback about the usability of the system. One of the main concerns was how to point the camera to the essential textual objects so that a meaningful interpretation is possible. As one of the volunteers stated, without earlier experience, it is hard to know what kind of texts could be available in the environment. One volunteer proposed an idea that the system should offer instantly some text excerpts from the current view that would help in the framing of the view with the camera. It was evident that if the framing is difficult, the users often try to perceive the space by touching, or just by taking repeated shots towards different directions.

Fig. 2. Exit from a subway station to a tram stop

Fig. 3. Entrance to a shop with the shop name clearly visible

The volunteers proposed several novel ways to use the system. One of them is a situation where a user arrives to a new environment and wants to know what kind of shops, products and discounts are available. Textual signs and information boards are considered useful especially since they are typically written in a clear manner. The volunteers found that suitable purposes for the use of the system are shopping and traveling. In shops one can locate products, examine their properties and thus compare them. In public transportation one can check timetables and traffic routes.

The volunteers had divergent opinions about the usability of the system in everyday life. One user postulated that it is much faster to ask help from a passer-by

than use the system. On the other hand, as seen in Section 2, some users note that they would not like to depend too much on the help from other people. One of the volunteers made an explicit point that he does not want to bother passers-by constantly. In addition, when alone at e.g. a bus stop, she might need to check the timetable independently. Another user feels that asking the names of the shops is tolerable but asking about advertisements or price comparisons is too intruding. This reason could encourage him to begin using the system although the speed of operation of the system might prove to be too slow after the initial excitement wears off.

The users point out that the usefulness of the system depends strongly on the easy portability and the capability to frame the view to be interpreted correctly. The users would appreciate the possibility to carry the system in a backpack and to use their hands only to take pictures. In addition, the framing of the views could be assisted by connecting it somehow to head movement. Also the procedures necessary to perform before the system is in operation raise some concern from the volunteers. To be truly usable, the procedure of setting up the system and starting it should be simple to do non-visually, as well as maintenance such as charging the batteries.

The results obtained by interpreting texts in a real-life surrounding environment with the system reflected the expectations of the volunteers. The accuracy of the interpretations was not high but yet often sufficient to give an overview of the textual content. Despite the limitations of the system, the volunteers considered the system to be generally useful since it adds to the independent initiative. One volunteer stated that he could begin to use even the rough prototype version right away in his everyday life.

6 Conclusions

The development of the system has positively shown that even with a quite modest level of technical expertise it is possible to create a useful computational solution for alleviating the problems of coping in everyday life. Existing devices and software can be harnessed to serve together in a novel way. From the perspective of software engineering, a truly open component-based architecture using existing modules enables to replace any component with a better component at any time. As long as copyright issues are taken sufficiently into account, this kind of product development can fruitfully support competition between manufacturers of different components. Due to the rapid rise of computing power and evolution in interoperability between components, today typical portable personal computers have the processing ability to carry on relatively demanding computational tasks.

To be useful, a device described in this article should be as portable as possible. In current research, the emphasis has been on mobile devices (see e.g. The Sypole Project [4]). The vOICe [10] project tries to convey visual imagery to aural information, and it has been implemented into a specific mobile phone (as The vOICe BEB), and is currently freely downloadable.

The vOICe BEB is a standalone application, so a natural direction for further development of portable text-to-speech is simply to use easy-to-acquire OCR and speech synthesis components, and integrate them into a mobile phone with a camera. In fact, some mobile phones will come already with an integrated speech synthesizer.

Simple OCR software could be used, and as the processing power in current mobile phones grows, better functionality can be achieved.

Components to build a working prototype into a mobile phone are already available. QuickTextScan from JSS Computing captures any text in the environment using the mobile phone's own camera and opens it for editing and passing to other applications [5]. Generic lightweight speech synthesis for mobile phones is also being developed (see e.g. VSpeak that provides a speech synthesizer working within the restrictions of contemporary mobile phones [18]).

Acknowledgments. The authors wish to thank the Finnish Federation of the Visually Impaired, people who replied to the survey, and especially those five persons that took part in the field trial. Moreover, the providers of software and hardware used in the prototype are gratefully acknowledged.

References

1. Technologies, C.A.: Inc. P2RD: Portable Print Reading Device (2004) http://www.catechnology.net/
2. Doermann, D., Liang, J., Li, H.: Progress in Camera-Based Document Image Analysis. In: ICDAR '03. Proc. seventh international conference on Document analysis and recognition, p. 606. IEEE Computer Society Press, Washington, DC, USA (2003)
3. Ezaki, N., Bulacu, M., Schomaker, L.: Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons. In: ICPR '04. Proc. Pattern Recognition, 17th international conference on (ICPR'04), vol. 2, pp. 683–686. IEEE Computer Society Press, Washington, DC, USA (2004)
4. Gaudissart, V., Ferreira, S., Mancas-Thillou, C., Gosselin, B.: Sypole: A Mobile Assistant for the Blind. In Proc. European Signal Processing Conference (EUSIPCO 2005), Antalya, Turkey (2005) http://tcts.fpms.ac.be/publications/papers/2005/eusipco05_vgsfcmbg.pdf
5. JSS Computing, Inc. QuickTextScan (2006) http://jsscomputing.com/quicktextscan/
6. Krishna, S., Little, G., Black, J., Panchanathan, S.: iCARE Interaction Assistant: A Wearable Face Recognition System for Individuals with Visual Impairments. In: Assets '05: Proc. 7th international ACM SIGACCESS conference on Computers and accessibility, pp. 216–217. ACM Press, New York, NY, USA (2005)
7. Krishna, S., Little, G., Black, J., Panchanathan, S.: A Wearable Face Recognition System for Individuals with Visual Impairments. In: Assets '05: Proc. 7th international ACM SIGACCESS conference on Computers and accessibility, pp. 106–113. ACM Press, New York, NY, USA (2005)
8. Lahti, L.: Computer-assisted acquisition of information for visually impaired (in Finnish). Master's thesis, University of Helsinki, Faculty of Science, Department of Computer Science, Report C-2006-32. (2006) http://ethesis.helsinki.fi/julkaisut/mat/tieto/pg/lahti/
9. Mallett, C.: AutoHotkey: Open Source Mouse and Keyboard Macro Program (2006) http://www.autohotkey.com/
10. Meijer, P.: The vOICe: A Synthetic Vision for the Blind (2006) http://www.seeingwithsound.com/
11. Mihara, Y., Sugimoto, A., Shibayama, E., Takahashi, S.: An Interactive Braille-Recognition System for the Visually Impaired Based on a Portable Camera. In: Proc. CHI'05: CHI '05 extended abstracts on Human factors in computing systems, pp. 1653–1656. ACM Press, New York, NY, USA (2005)

12. Nagel, H.-H.: Steps toward a Cognitive Vision System. AI Magazine 25(2), 31–50 (2004)
13. Nullsoft Ltd. Winamp: Media Player for Windows (2006) http://www.winamp.com/
14. Paradise, J., Mynatt, E.D., Williams, C., Goldthwaite, J.: Designing a Cognitive Aid for the Home: A Case-Study Approach. In: Assets '04: Proc. 6th international ACM SIGACCESS conference on Computers and accessibility, pp. 140–146. ACM Press, New York, NY, USA (2004)
15. Ross, D.A., Lightman, A.: Talking Braille: A Wireless Ubiquitous Computing Network for Orientation and Wayfinding. In: Assets '05: Proc. 7th international ACM SIGACCESS conference on Computers and accessibility, pp. 98–105. ACM Press, New York, NY, USA (2005)
16. Timehouse Oy. Mikropuhe: Finnish Speech Synthesis Software (2003) http://www.mikropuhe.com/mikropuhe.asp
17. TopSoft Ltd. TopOCR: Optical Character Recognition Software (2005) http://www.topocr.com/
18. VoiceSignal Technologies, Inc. VSpeak: Speech Synthesis for Mobile Phones (2006) http://www.voicesignal.com/solutions/applications.php3
19. Yang, J., Gao, J., Zhang, Y., Waibel, A.: Towards Automatic Sign Translation. In: HLT '01: Proc. 1st international conference on Human language technology research, pp. 1–6. Association for Computational Linguistics, Morristown, NJ, USA (2001)
20. Zandifar, A., Chahine, A.: A Video Based Interface to Textual Information for the Visually Impaired. In: ICMI '02: Proc. 4th IEEE international conference on Multimodal interfaces, p. 325. IEEE Computer Society, Washington, DC, USA (2002)