Read It Aloud to Me

Sergio Celaschi[1(✉)], Mauricio Sol Castro[2], and Sidney Pinto da Cunha[1]

[1] Centro de Tecnologia da Informação Renato Archer, Campinas, Brazil
{sergio.celaschi,sidney.cunha}@cti.gov.br

[2] Fundação de Apoio à Capacitação em Tecnologia da Informação,
Rod. D.Pedro I, km 143,6, Campinas, SP 13069-901, Brazil
https://www.cti.gov.br

Abstract. The universal design applied to assistive technologies can help visually impaired person perform some day-to-day tasks as well as everybody. With this aim, the present work focuses on development of photo-to-speech instruments for the visually impaired person. It allows the user to hear text typed on a sheet of paper or written/posted on a wall. To achieve that aim a set of image capture and processing frameworks such as Optical Character Recognition (OCR) and Text to Speech Synthesis (TTS) were integrated. The first versions of the OCR based speech synthesis systems were developed for our native language, Portuguese. A preliminary desktop version was designed under Windows OS, and a version for mobile devices was developed as an application for Android devices. In this paper, we summarize efforts to develop and test a desktop and a mobile version of autonomous photo-to-speech instruments for the visually impaired. The project consisted of integration of selected components, and the CPU applications governing several functionalities: capture of images by the CCD camera; image preprocessing; OCR framework for text recognition; and finally the process of TTS, producing a synthesized voice.

Keywords: Assistive technology · Text reading · Speech synthesis · OCR · Photo-to-speech · Blind · Visually impaired · Universal design

1 Introduction

Many software and applications are especially dedicated to visually impaired person, some that improve accessibility like Talkback (Android) and Voice Over (IOS), and screen readers. Others example are: ZoomReader (iOS) [1], o CapturaTalk [2], Google Translator [3], LookTel Money Reader [4] (Money reader-iOS), TapTapSee (iOS) [5] that identify objects to blind person out loud, a Brazilian Portuguese TTS Alcance-CPqD [6], helping blind person to access a lot of smart phone services. See the appendix A, for a list of applications. The present paper describes the development of photo-to-speech instruments for visually impaired persons [7] with the following prerequisites: to be free of charge to users in general, easy to install and easy to use, available locally in the country, useful and very understandable in Brazilian Portuguese, locally processed

M. Antona and C. Stephanidis (Eds.): UAHCI 2017, Part II, LNCS 10278, pp. 260–268, 2017.
DOI: 10.1007/978-3-319-58703-5_19

without internet and cloud connections, low memory space requirements and compatible with most common used operational systems such as windows, IOS and Android. Specifically, this engine gives access to printed information, leading images of printed texts to audio voice. It allows the user to hear text typed on a sheet of paper or written/posted on a wall, in outdoors and billboards. The application has the following sequence: capture of images by the CCD camera of A4 printed text, preprocess the image, optical character recognition with an OCR - Optical Character Recognition software [8], extraction of the text [12], and synthesized voice generation with a TTS-text-to-speech application.

2 Used Technologies

2.1 Image Acquisition

The specifications of the digital camera are mandatory, which require mobile devices with 6–8 Megapixels resolution, main camera to capture images from an A4 size paper. The quality and resolution of the image are crucial for the preprocessing stage and character recognition. The OCR efficiency requires roughly up to 10 × 10 pixels per character. The preprocessing stage improves the overall performance of the text recognition providing feedback to the OCR framework. Finally the TTS framework provides the audio stream to the internal Digital Signal Processing-DSP unit to drive the speakers. The text classification is a binary output in which an input text image is considered readable or non-readable without any character recognition. The OCR will be described in the next subsection. Figure 1 shows a flowchart of steps in the process, as taking a picture of the document, preprocessing as binarization, some small corrections in the skews, submitting to the OCR, calling the voice synthesizer, and the output to user.

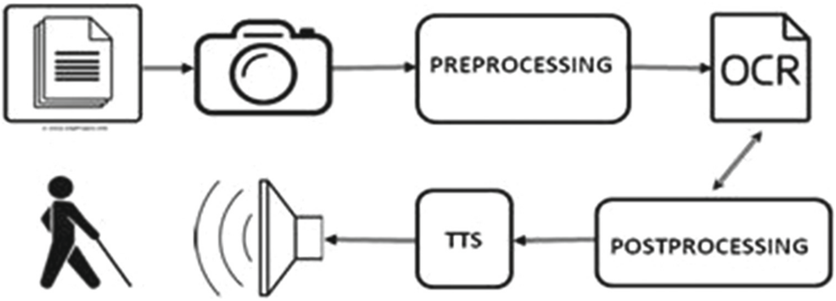


Fig. 1. Flowchart of software operation.

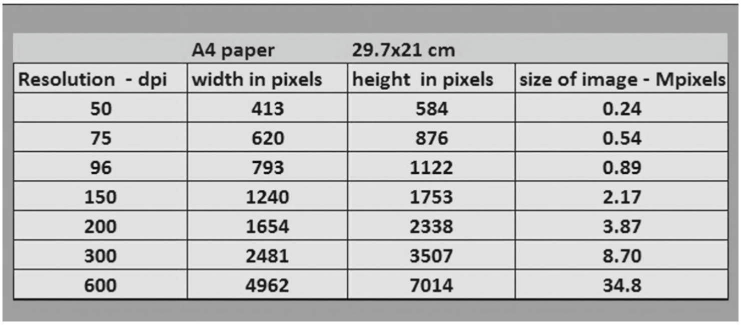
2.2 Optical Character Recognition Tesseract

The Tesseract technology of Optical Character Recognition (OCR) does part of the task. It enables the extraction of texts from image data. Tesseract began as a PhD research project in HP Labs, Bristol. Today, Tesseract is an open-source OCR engine distributed by Google Inc. This technology has been used in photographed documents which one can play and listen to the content in audio formats, using TTS - resembling the spoken form of the same text, as read by a native. Well established in the field of computer science, early approaches to language research focused on automating the analysis of the linguistic structure of language. This research also relied on developing basic technologies such as fast digital processors, multi megapixel CCD cameras, OCR, machine translation, speech synthesis, among others. Nowadays, such tools are employed in real-world applications, creating spoken dialogue systems and speech-to-speech translation engines. Google Translate - GT is one such a practical application. It is a multilingual machine for text translation, speech, or real-time text images, from one language into another. It offers a web interface, and cloud interfaces for Android and IOS mobiles. It has one requirement: the photo-to-speech GT functionality is not autonomous, but depends on cloud processing. To improve system quality performance [9], a preprocessing stage of the image was carried out before submitting it to character recognition. The preprocessing stage improves the overall performance of the text recognition providing feedback to the OCR framework. The preprocessing include the following sequence of operations: color to gray scale transformation - 8 bits/pixel, image "binarization" (1 bit/pixel), image rotation and median filtering. Those image operations used OpenCV functions library.

OpenCV [10] written in optimized C/C++, with multi-core processing and real-time applications is very useful for our software. EmguCV [11] was used as a cross platform to .NET framework to call functions of OpenCV, to capture image of the printed text to be recognized. The Ziggi HD camera employed with 3264 × 2448 pixels, has resolution enough to recognize the characters. The captured images needs to be rotated to be in portrait position. It was observed that the OCR needs a resolution of 300 dpi(dot per inch) approximately or a 10 × 10 pixels per character to have good results in the recognition.

Some image operation are done inside Tesseract. The implementation and optimization of this autonomous photo-to-speech instrument aims, in a short period of time, the design of a fully accessible equipment. Text reading, available in digital format, for the visually impaired, requires the text conversion to the Braille reading system or, more recently, a digital speech synthesizer. Nowadays, most published printed works does not include audio versions nor Braille reading. Thus, the development of an autonomous and portable machine that captures images containing texts, converting them into speech [13] is greatly useful for visually impaired person [14].

Figure 2 shows relations, for an A4 paper, of resolutions, width, height and size of images. In our case it is necessary resolutions about 300 dpi, for better recognitions



	A4 paper	29.7x21 cm	
Resolution - dpi	width in pixels	height in pixels	size of image - Mpixels
50	413	584	0.24
75	620	876	0.54
96	793	1122	0.89
150	1240	1753	2.17
200	1654	2338	3.87
300	2481	3507	8.70
600	4962	7014	34.8

Fig. 2. Size of A4 and the image resolution, comparisons.

2.3 Voice Synthesizer

The software application can use the embedded voice synthesizer [15] of the operational system or could use such free screen reader such as NonVisual Desktop Access-NVDA, eSpeak, or Alcance Voice Synthesizer, a Brazilian initiative project including voice synthesizer.

3 Operation and Features

Aiming to be very easy and friendly to use, the app needs few commands to be operated, as seen by the list of commands below. The accessibility of the desktop application version relies fully on the keyboard as follows:

- (1) Return/Enter if the program is in idle mode, triggers a frame capture, and the whole cycle shown in Fig. 1.
- (2) Esc if the program is in idle mode, closes the application - if a speech is playing, it is canceled.
- (3) Left/Right keys decrease/increase speech rate.
- (4) Space pause/resume a speech.
- (5) Bar language choice.

4 Architecture

Figure 3 shows an operation with a stand to fix camera. So, better pictures of the texts can be obtained. In close up photos is necessary to avoid camera shakes. The Fig. 4 below shows the state machine model for a desktop software [17]. Using version for mobiles, to take close-up photos of text, is better to use a stand as shown in figure below. Figure 5 shows a special stand with four legs, designed for mobiles. It has the right distance for picture a A4 text, and the legs could guide the position of the paper.

For example, the accessibility of the mobile previous version operates under the native voice assistant Samsung. One touch on the screen triggers a frame

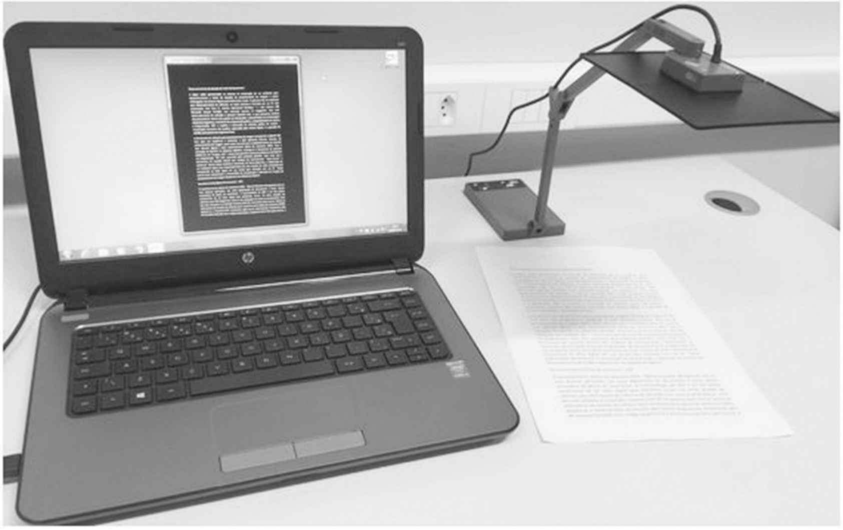


Fig. 3. Desktop configuration, with a camera and adjustable stand.

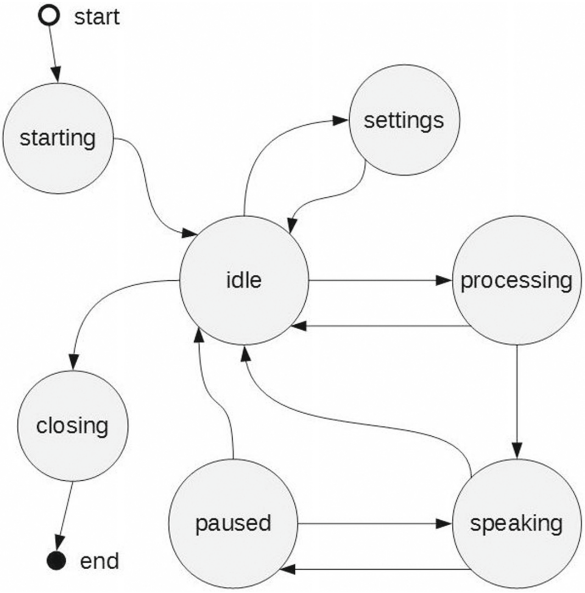


Fig. 4. Architecture - state machine model.

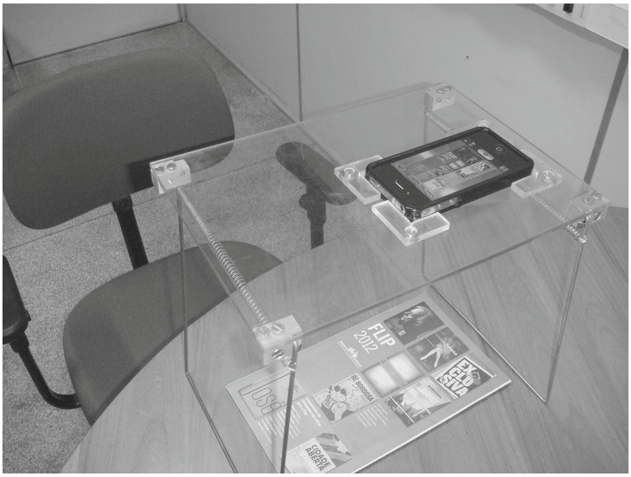


Fig. 5. FourPod - stand for mobile capture - patent pending.

capture, and the whole cycle. The "LETEX" app applies small corrections misalignment for angular position of text columns, up to 30°. The text could be positioned upside-down.

5 Blind Users Review and Performance

5.1 Review

According a reviewer a digital reader like this one has good impact over the users' lives, as in the case of independence and privacy when used with head phones, for example. Independence because the user can digitalize a text or a document alone without any help. The program is easy to open and run, and there are few buttons to push. Also the alignment of the paper is guided by a frame. The user can choose the speed and pitch of the voice synthesizer. Compared to equivalent commercial equipments this one is affordable, using a personal computer and/or mobile devices. In the case of mobile devices it is more useful in reading written characters in signs and outdoors. Although is not so easy to capture sharp photos through mobile caused by the vibrations. So, is better to use a stand as shown to have good text pictures.

5.2 Performance

Summarizing the pros and cons of performance pointed out by users and also observed in the preliminary tests are the following: Fig. 6 shows a two-columns text, that was inverted to negative image, and could be recognized by Tesseract OCR.

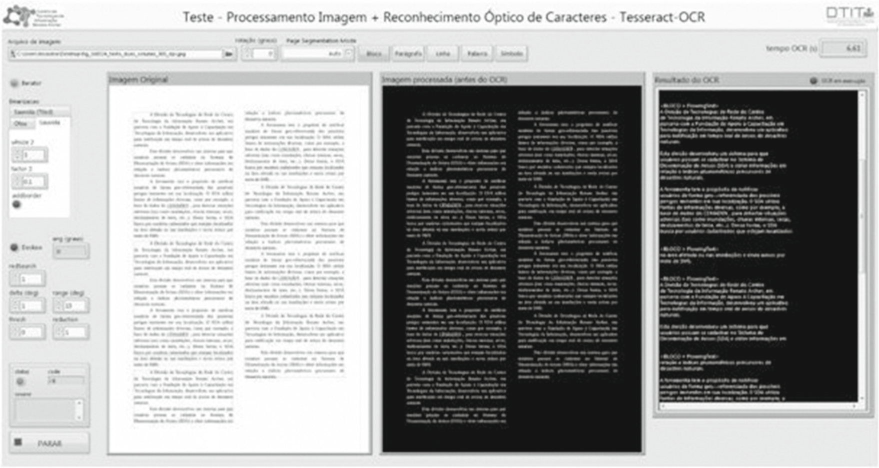
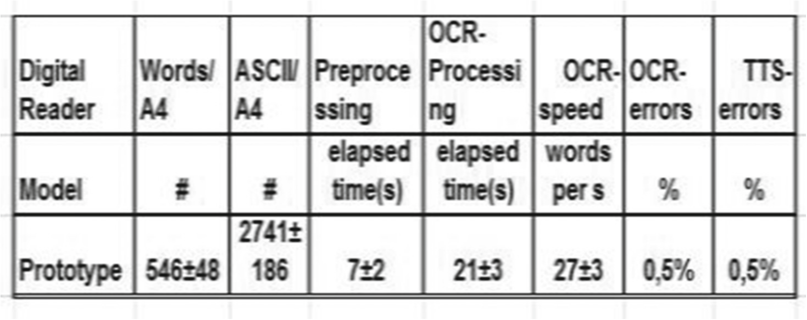


Fig. 6. Two-columns text, that also can be recognized by Tesseract-OCR.

Figure 7 shows results of tests, with a very few errors, and a short elapsed time in the processing, indicating that the project can succeed, and the software is useful for blind persons to use. Following the premises of simplicity, few commands, affordability, few equipment requirements.



Digital Reader	Words/ A4	ASCII/ A4	Preprocessing	OCR-Processing	OCR-speed	OCR-errors	TTS-errors
Model	#	#	elapsed time(s)	elapsed time(s)	words per s	%	%
Prototype	546±48	2741± 186	7±2	21±3	27±3	0,5%	0,5%

Fig. 7. Figure shows some data about performance of the reader.

5.3 Pros

- The software is a direct access to a document after the picture shot.
- With a guide platform it is very easy to align a document to be photographed.
- Its a very simple and intuitive to use and quite cheap to acquire.
- Desktop version is very useful for documents reading and the mobiles version is better to read outdoors, indoor building directions and street signs.
- User can choose the speed of speech and the voice pitch.

5.4 Cons

- It will be practical if users could save the file (not implemented yet).
- It will be useful if the program warns the users about impossibilities and causes of not reading certain documents or texts (not implemented yet).
- Blind users have difficulties in framing an A4 printed paper with a smart phone/mobile.
- Vibrations in the smartphone cause bad images and difficulties in the OCR.
- The mobile version is very useful for outdoors sign reading as compared to document reading.

6 Conclusion

The proof of concept has shown feasibility of the project [18]. It has practical importance mainly for visually impaired persons to access printed matter in general. The independence of connection with internet implies no money waste with data plans for smartphones. The software and application are friendly and very easy to use, with few commands. The version for desktop is better for reading printed texts in A4 format and the version for mobiles is useful for street signs and outdoors.

7 Future Work

The results of the preliminary work with this project has shown the possibility of adding other functionalities like saving the read text in any desired format, even in audio formats. Without enhancing complexity, integrating the Brazilian voice synthesizer, Alcance-CPqD, as the default choice, and voice warns, like “there is no text!”, “please realign the paper!”.

8 Final Comments

By the end of 2011, the Brazilian federal government created a plan for people with disability. The plan entitled “Living without Limits” [16] has four main branch areas: access to education, health care, social inclusion and accessibility. It involves cooperation of 15 federal agencies, states and municipalities. This project has been supported by one of these Brazilian federal agencies - FINEP, under contract number 01.13.0038.00 coordinated by Funda de Apoio Capacitao a Tecnologia da Informao - Facti. Shown in caption of Fig. 5, the FourPod - stand for mobile capture has a patent pending [19].

Acknowledgments. We have to thank FINEP for the financial support of the project and CTI Renato Archer for providing the supporting infrastructure.

A Appendix

It is possible to check on the internet the available commercial products. Check the addresses below:

- Sara CE - http://www.freedomscientific.com/Products/LowVision/SARA
- Eye-pal Solo - http://www.freedomscientific.com/Products/Blindness/EyePalSOLO
- Poet Reading Machine - http://www.baum.de/cms/en/poetcompact2/
- LavoiceSolo Reading - https://www.maxiaids.com/
- Zoo-Ex - http://www.abisee.com/Zoom-EX.html
- Magnilink Voice - http://lviamerica.com/products/readingmachine/magnilink-voice

References

1. ZoomText, ZoomReader. http://www.zoomtext.com/products/zoomreader/
2. iansyst Ltd.: CapturaTalk. http://www.capturatalk.com/
3. Google Inc.: Google Tradutor. https://play.google.com/store/apps/details?id=comgoogle.android.apps.translatehl=ptBR
4. LookTel: LookTel Money Reader. http://www.looktel.com/moneyreader
5. Image Searcher Inc.: TapTapSee. http://www.taptapseeapp.com/
6. https://www.cpqd.com.br/solucoes/cpqd-alcance/
7. Netoa, R., Fonseca, N.: Camera reading for visually impaired people. Procedia Technol. **16**, 1200–1209 (2014). In: International Conference on Health and Social Care Information Systems and Technologies - HCIST 2014
8. Smith, R.W.: The extraction and recognition of text from multimedia document images. Ph.D. Thesis, University of Bristol, Bristol (1987)
9. Mithe, R., Indalkar, S., Divekar, N.: Optical character recognition. Int. J. Recent Technol. Eng. (IJRTE) **2**(1), 72–75 (2013). ISSN: 2277-3878
10. Itseez: OpenCV (Open Source Computer Vision). http://www.opencv.org
11. Emgu: Emgu CV. http://www.emgu.com/wiki/index.php/Main
12. GitHub: Tesseract Open source OCR Engine (main repository). https://github.com/tesseract-ocr/tesseract
13. Hirschberg, J., Christopher D.M.: Advances in natural language processing. Sci. Mag. **349**(6245), 261–266 (2015)
14. Rodrigues, A., Montague, K., Nicolau, H., Guerreiro, T.: Gettingsmartphones to talkback: understanding the smartphone adoption process of visually impaired users. In: Proceedings of the 17th International ACM SIGACCESS Conference, ASSETS 2015, Lisbon, pp. 23–32 (2015)
15. Dutoit, T.A.: Short Introduction to Text-to-Speech. Kluwer Academic Publishers, Dordrecht, Boston, London (1997)
16. http://zeroproject.org/policy/brazils-billion-dollar-national-plan-for-inclusive-education/
17. Singla, S.K., Yadav, R.K.: Optical character recognition based speech synthesis system using LabVIEW. J. Appl. Res. Technol. **12**(5), 919–926 (2014)
18. Leitor Digital. https://www.youtube.com/watch?v=SEM6slss2ls
19. Instituto Nacional da Propriedade Industrial, Depsito de pedido nacional de Patente, pedido BR 10 2013 0295515 A2. http://www.inpi.gov.br/