Implementation of a Speech Based Interface System for Visually Impaired Persons

Adil Farooq, Ahmad Khalil Khan, Gulistan Raja

Department of Electrical Engineering, UET Taxila, Pakistan
ahmad.khalil@uettaxila.edu.pk

Abstract: Text to speech system (TTS) has become more popular to visually impaired people to assist them on reading textual information from digital sources and printed resources such as image acquired texts from newspaper and magazines. This paper describes the implementation of a voice based human-computer interface system with a complete text recognition and speech processing capability. A new idea is proposed to assist the visually impaired with heading read and skip concept that allows free maneuvering capability making it much more efficient than the existing readers. The method uses windows text to speech conversion and image recognition (Optical Character Recognition) technique to analyze and extract textual information from digital scanned images. Our research uses an open source engine Asprise OCR for text extraction and is expressed in audible format. The implementation is done in Microsoft visual studio using C sharp. The results show that our system's response is best for image with dimension 1557x2272 and resolution of 200 dpi in terms of processing speed and correctness. The time utilized from image capture to final speech output was approximately 2.479 seconds.

[Adil Farooq, Ahmad Khalil Khan, Gulistan Raja. **Implementation of a Speech Based Interface System for Visually Impaired Persons.** *Life Sci J* 2013;10(9s):398-400]. (ISSN: 1097-8135). http://www.lifescience site.com 57

Keywords: OCR, TTS, Open source, Speech synthesis

1. Introduction

Text-to-speech (TTS) conversion plays a vital role in our daily life applications such as reading online e-mails and fax, user interactions via dialogue systems, automatically reading text, telephonic automated systems, multimedia, GUI's etc (Rehman and Saba, 2011). In general, TTS can generate synthesized voice without limit according to the input textual information. Natural and fluent speech is the most important issue for the development of TTS (Yeh and Hwang 2005). The portable aided systems for visually impaired persons can perform image to text recognition or text to speech conversion functions separately. These systems can significantly improve the level of education and quality of life for visually impaired men and women. However, most of the handheld devices and software currently available for this purpose are quite expensive. These devices cannot be easily expanded and are based on closed systems. Moreover, little effort is made to achieve integration between image to text and text to speech conversion systems (Battaglia 2012). Image to text and text to speech conversion systems can be integrated together by using an Application Programming Interface (API) of third party. Enhancement and modification of an open source software can be developed, which may result in more effective and cheaper products (Hedgpeth and Black 2006).

Researchers have done a lot of work on text to speech technology. However, reading assistance for disabled persons is one of the most explored area of research: Upson (2007) suggested that hardware aimed at assisting visually impaired people should not cost more than $1000 to gain large diffusion. Shinohara and Tenenberg (2009) discussed the needs of a visually impaired user and the way technology can aid them in everyday life (Rehman and Saba,2012).

Watanabe (2001) applied artificial neural networks extensively to document analysis and recognition for image to text conversion. Marinai et al (2006) used the word-level indexing of modern printed documents which is difficult to recognize using current OCR engines. By means of word-level indexing, it is possible to retrieve the position of words in a document. Song et al (2005) explained a merged character segmentation and recognition method based on forepart prediction, matching and character masking. One of the difficulties faced in OCR systems is that the shading artifacts commonly occur. Meng et al (2008) proposed a method for document shading correction using local and adaptive binarization techniques to produce satisfactory text output.

It is observed from literature review that researchers have performed extensive researches on text to speech conversion and image to text recognition techniques separately. The integration of these techniques into one platform has not yet been significantly explored (Saba and Rehman,2012).

This paper describes the combined implementation of image to text and text to speech technique using OCR and speech synthesis for visually impaired persons. The rest of the paper is organized as follows. First we describe the implementation of our speech system. This is followed by experimental results, discussion and finally the conclusion.

2. Implementation of Speech Based System

The block diagram of proposed implementation for combined image to text and text to speech system is shown in Figure 1.

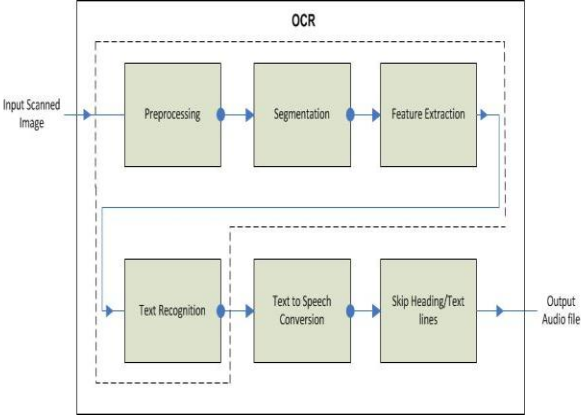


Figure 1. Proposed Block Diagram for Speech Based Interface System

The input to system is the digital scanned textual image which is preprocessed for image enhancement, noise removal, image thresholding and skew detection/correction. The preprocessed digital image is then partitioned into multiple fractions during segmentation process. These fractions formed the local zones of the image which was useful to extract features for character recognition and assigned a label to each pixel in an image. The segmented image is used to extract a set of features and maximize the recognition rate in feature extraction. The methods used in feature extraction are zoning and projection histograms. In zoning, the frame containing the character is divided into overlapping and non-overlapping regions to calculate densities of object pixels in the regions. It is calculated by finding the number of object pixels in each region divided by total number of pixels. The numbers of pixels in horizontal, vertical, left and right diagonal directions were counted by the projection histogram. The editable textual data is obtained during text recognition stage. The textual data is sent to speech synthesizer which used text to speech function of the windows operating system. Once the textual information is added to the Text to Speech conversion module, the user can listen the information from the GUI by selecting the speak button option. A new idea is proposed to ease the maneuvering of reading textual information, using heading read and skip concept. Whenever the user wishes to skip a heading topic, s/he on listening the heading, selects the skip heading option from the select buttons. The system automatically skips to the next heading to continue.

The pseudocode for heading read and skip is as follows:

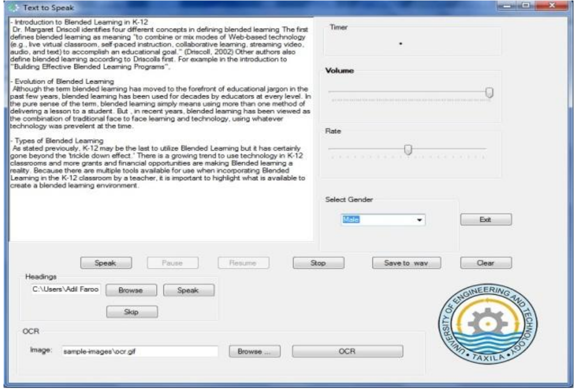
```
do  Text-to-Speech Conversion (Text Data)
start
    while  Text = Heading
        if   Heading is relevant
             Continue reading
        else
             Jump to next heading
end do
```

When the textual data is read via speech synthesizer, the information can also be saved by the user in audible format (.wav file). All the above mentioned selected options are heard in the developed GUI using single tab button from the system keyboard. The visually impaired person with little practice can easily be trained to this system.

3. Results and Discussion

Our implemented speech system described above is capable to read textual information from scanned images and generate output speech. The system is designed in Microsoft Visual Studio 2010 simulation environment using C# language. Open source API for OCR used is from Asprise OCR. Handheld computer with windows 7 operating system, 2.10 GHz processor, 2GB RAM, 200GB hard disk is employed for experimentation.

The complete developed GUI of our systems with select button options is shown below in Fig 2.



- Introduction to Blended Learning in K-12
Dr. Margaret Driscoll identifies four different concepts in defining blended learning The first defines blended learning as meaning "to combine or mix modes of Web-based technology (e.g., live virtual classroom, self-paced instruction, collaborative learning, streaming video, audio, and text) to accomplish an educational goal." (Driscoll, 2002) Other authors also define blended learning according to Driscolls first. For example in the introduction to "Building Effective Blended Learning Programs".

- Evolution of Blended Learning
Although the term blended learning has moved to the forefront of educational jargon in the past few years, blended learning has been used for decades by educators at every level. In the pure sense of the term, blended learning simply means using more than one method of delivering a lesson to a student. But , in recent years, blended learning has been viewed as the combination of traditional face to face learning and technology, using whatever technology was prevelent at the time.

- Types of Blended Learning
As stated previously, K-12 may be the last to utilize Blended Learning but it has certainly gone beyond the 'trickle down effect.' There is a growing trend to use technology in K-12 classrooms and more grants and financial opportunities are making Blended learning a reality. Because there are multiple tools available for use when incorporating Blended Learning in the K-12 classroom by a teacher, it is important to highlight what is available to create a blended learning environment.

Select Gender: Male

Headings: C:\Users\Adil Faroo

OCR

Image: sample-images\ocr.gif

Figure 2. Text to Speech GUI

Different configuration variables like volume, rate and gender can be set by the user. Image file is selected from browse option and converted to text using OCR button. Editable text file can also be given as input using Text Browse option. Speech output can be taken from Speak button and saved into audio file by selecting save to wav.

Table 1 shows the average time for text to speech conversion. The textual data used in proposed research were taken from paper written by Yurtay et al (Yurtay et al 2011).

Table 1. Text time result for TTS Module

Text No.	No. of words in Text	Time taken (secs)		
		Previous Rate	Our Rate (low)	Our Rate (Normal)
1	1	0.850341	2.195	1.83
2	2	1.290039	3.210	2.123
3	4	3.228515	4.312	2.182
4	12	5.57373	5.413	3.238
5	80	39.115234	34.3398	20.1968
6	71	47.38208	41.4076	24.2634
7	180	80.46997	81.8034	47.4649

The total words in given text are synthesized with varying pitch rates. The results show that our system took less time for processing speech output as the number of words in the text was increased.

Table 2 below shows the result of text line/heading skip concept.

Table 2. Average Saved of TTS System

No. of Lines	**No. of words**	**Time taken (secs)**		
		Low Rate	Medium Rate	High Rate
1	12	10.946	6.503	3.253
2	24	19.181	10.023	7.625

Random textual data having 12 words per line is taken and average time calculated. It is observed for medium rate on skipping 1 line, time taken was 6.503 seconds while for 2 lines was 10.023 seconds. Similarly for multiple lines the average time was calculated by doubling the word lines for various pitch rates. For skipping 3 lines per heading the average time saved was 6.503+10.023=16.526 seconds. In this way, the heading skip feature of our system was not only used to skip unnecessary information but also saved a lot of time. The result of image to text conversion is shown below in Table 3.

Table 3. Text detection for 23 line image using OCR

Image (dpi)	Conversion Time (sec)	Correct Detected	Wrong Detected	Precision (%)
100	1.034	2	21	8.6
200	2.479	19	4	82.6
300	5.208	10	13	43.4

Precision = CorrectDetected / (CorrectDetected + WrongDetected)

An image having text word length of 23 lines is taken with different resolutions (dpi). The result in terms of processing speed and correctness is calculated. It shows that the system's response is the best for image with dimension 1557x2272 and resolution of 200 dpi. The time utilized from image capture to final speech output is approximately 2.479 seconds with maximum precision of 82.6%.

4. Conclusion

We have described a complete Image to Text and Text to Speech conversion system for reading printed and digital documents. Experimental results show the best result for 200 dpi image resolution with almost 2 second conversion time and precision of 82.6%. The system can be used to aid visually impaired persons for reading telephonic messages, online e-mails, fax and printed documents.

References

[1] Yeh CY and Hwang SH (2005) Efficient text analyser with prosody generator-driven approach for Mandarin text-to-speech. IEE Proceedings of Vision, Image and Signal Processing. 152: 793–799.

[2] Battaglia F (2012) An Open Architecture to Develop a Handheld Device for Helping Visually Impaired People. IEEE Transaction on Consumer Electronics**.** 58:1086-1096.

[3] Hedgpeth T and Black JA (2006) A demonstration of the iCARE portable reader. Proceedings of the 8th international ACM SIGACCESS conference on computers and accessibility, Portland, 279-280.

[4] Upson S (2007) Tongue vision: a fuzzy outlook for an unpalatable technology. IEEE Spectrum. 44:44-45.

[5] Shinohara K and Tenenberg J (2009) A blind person's interactions with technology. Communications of the ACM. 52:58-66.

[6] Saba, T. and Rehman, A. (2012). Effects of Artificially Intelligent Tools on Pattern Recognition, International Journal of Machine Learning and Cybernetics, 4(2),155-162.

[7] Watanabe T (2001) Merits of open-source resolution to resolve a digital divide in information technology. Proceedings of the First International Conference on The Human Society and the Internet - Internet Related Socio-Economic Issues, Seul, Korea, 92-99.

[8] Marinai S, Marino E, and Soda G (2006) Font Adaptive Word Indexing of Modern Printed Documents. IEEE Transactions on Pattern Analysis and Machine Intelligence. 28:1187-1199.

[9] Rehman, A. and Saba, T. (2012) Neural Network for Document Image Preprocessing" Artificial Intelligence Review, DOI: 10.1007/s10462-012-9337-z.

[10] Song J, Li Z, Lyu MR and Cai S (2005) Recognition of Merged Characters Based on Forepart Prediction, Necessity-Sufficiency Matching, and Character-Adaptive Masking. IEEE Transactions on Systems, Man, & Cybernetics.35:2-35.

[11] Meng G, Zheng N, Du S, Song Y, and Zhang Y (2008) Shading Extraction and Correction for Scanned Book Images. IEEE Signal Processing Letters. 15:849-852.

[12] Yurtay, Bicil N, Celebi Y, Cit S, Dural G, Deniz (2011) Library Automation Design for Visually Impaired People. Turkish Online Journal of Educational technology – TOJET. 10:255-260.

[13] Rehman, A. and Saba, T. (2011). Performance Analysis of Segmentation Approach for Cursive Handwritten Word Recognition on Benchmark Database. Digital Signal Processing, 21(3), pp. 486-490.