Published in Proc. of International Conference on Advance Computer Science & Information System (ICACSIS) 2010, pp.393-396, Bali, Indonesia, 2010

TextReader: An Intelligent System for People with Visual Impairment

Anto Satriyo Nugroho, Irfan Suwadi, Vitria Pragesjvara, Ratih Irbandini, Oskar Riandi, Made Gunawan, Dwi Handoko
Center for Information & Communication Technology, Agency for Assessment & Application of Technology
BPPT 2nd building 4F, Jalan M.H.Thamrin 8 Jakarta, Indonesia 10340
Email: asnugroho@inn.bppt.go.id

***Abstract*— Visual impairment is a vision loss caused by trauma, disease, and congenital or degenerative conditions. This disability condition makes visual impairment people difficult or even could not understand the textual information at all. The objective of the research is to develop "TextReader", an intelligence system to help people with visual impairment in Indonesia to increase their accessibility on textual information. "TextReader" combines Optical Character Recognition (OCR) and Text–to-speech Synthesizer (TTS) which converts image input – as scanning result of the textual document- into character symbols, then from character symbols into voice. All the system were developed based on Free Open Source Software policy, thus can be used by the visual impaired at no cost. Preliminary experiments demonstrated the performance of "TextReader" by perfectly recognized two documents in Bahasa Indonesia and converted the result into audio format.**

***Keywords*—OCR, TTS, Visual impairment, Bahasa, Open source**

I. INTRODUCTION

Visual impairment is one of national problem in Indonesia with the ratio 1.5% of the population as reported in the survey conducted by the Ministry of Health in 1993-1996. People belongs to this category has very limited access to the information. Braille characters could be considered as the only way to access the textual information. However, only few of the literatures are provided in Braille thus their accessibility to the information is still very limited. To help these people, various research and development of supporting systems have been conducted in modern contries [1]. Examples of such equipments are Optical Braille Character Reader (OBR), virtual sound based 3D information visualization [2]. This system is however still relatively expensive, and not accessible for the visual impaired in Indonesia. Nevertheless, only few activities found in Indonesia to help the people with visual impairment. One of such efforts is e-book reading invitation which is initiated by Mitra Foundation [3] but yet still not optimal since it requires great effort and involving large number of people. This situation motivated us to develop an intelligent system to improve the accessibility to the textual information of the visually impaired in low cost and accessible by the communicty in Indonesia. "TextReader" is the system developed in this study by combining a self-customized Optical Character Recognition (OCR) and Text To Speech Synthesizer (TTS), thus converting the information from analog to digital image by OCR, then it is converted into voice signal by Indonesian TTS. With TextReader system, in the future the visual impaired people will not depend on the Braille anymore to access the information written in book, newspapers, and even the text in outdoor environment.

This paper is organized as follows: Sec.2 described the architecture of "TextReader" system, Sec.3 reported the experimental results which will be summarized and concluded in Sec.4.

II. SYSTEM ARCHITECTURE

"Text reader" consists of five central component: (i) pre-processing, (ii) segmentation, (iii) character classification, (iv) post-processing and (v) text to speech synthesizer, as shown in Fig.1.

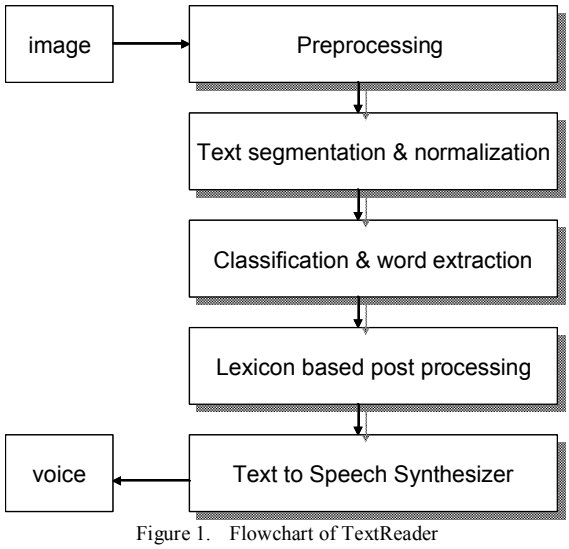

Figure 1. Flowchart of TextReader

(i) Preprocessing
Preprocessing is conducted by a series of standard image processing procedures to remove the noise, image enhancement making the image easier to be processed by the subsequents modules.

(ii) Text Segmentation & Normalization
Text Segmentation aims to extract text area of an image, removing the non text area. This module is simply based on histogram analysis of the intensity of the pixels. The first step is analyzing the image intensiy horizontally, find the line area from the histogram. Assumed that the text is using common font size, the line will have almost similar height. Image has distinguished histogram characteristics compared to the line, that it has wider range and deviates from the average height of the lines. Once line has been extracted, the same histogram is generated over the line, thus each of the character is extracted individually. In this study, we assumed that the text is written in Indonesian that uses Roman alphabet. Roman alphabet is constructed by single segment and its segmentation is much easier compared to multi-segment type such as Japanese Kanji characters, sinch which required complex processing [4]. However, some character fonts is not easy to segment, due to the variety of its width and in some cases to adjacent characters are connected. Here we used Courier font that has occupies the same amount of horizontal space, thus the problem is simplified and the study will be focused on the development of the system prototype. After the character is extracted, it will be normalized into 12x12 mesh, creating a 144-dimensional vector representation of the characters. The normalized image will be used as a feature in classification module

(iii) Character classification and word extraction
Classification module is implemented by Multilayer Perceptron Neural Network which is trained by Backpropagation algorithm. The neural network has three layers: input, hidden and output layer. Each layer has 144, 50 and 36 neurons respectively. The output of the classification is 26 Roman alphabets and 0-9 numeral characters. The character result of classification process will be grouped into a word by detecting a white space in the original image, that significantly exceeded the average character width.

(iv) Lexicon based Post processing
The goal of lexicon based post processing component is to correct the output word from classification component by referring to a dictionary. The word searching process is implemented using linear search, and the words similarity is measured using Longest Common Subsequences (LCS) algorithm [5]. LCS is a dynamic algorithm which used to find similarity and dissimilarity between two or more string. LCS is commonly used to find similarity and a consensus among DNA sequences. The pseudocode of the algorithm is depicted in Fig.2.

1. for i ← 0 to n
2. Si,0 ← 0
3. for j ← 1 to m
4. s0,j ← 0
5. for i ← 1 to n
6. for j ← 1 to m
7. $$s_{i,j} \leftarrow \max \begin{cases} s_{i-1,j} \\ s_{i,j-1} \\ s_{i-1,j-1} \quad if \quad v_i = w_j \end{cases}$$
8. return $(s_{n,m}, b)$

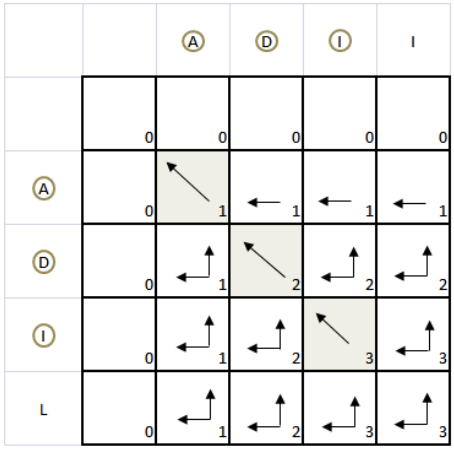

Figure 2 Pseudocode of Longest Common Subsequences

First, fill in the entire column and row 1 with the number 0 (zero) because the row and the column did not related to the same character. Furthermore, to resolve column 2 row 2, compare the figure in column 1 row 1, column 2 row 1, column 1 and row 2. The greatest number among those three celss are selected and written into column 2 row 2. But if the corresponding column and row have the same character, the number in the cells is incremented. The same process is conducted throughout the rest cells. The similarity between the two strings is found in the bottom right corner cell. The score is then processed by the following equation:

$$score = \frac{LCSoutput \times recognized\_word\_length}{corrected\_word\_length} \quad (1)$$

The corrected character is then converted into phoneme code consists of phoneme, duration, and pitch, e.g.

V 51 25 114

The string above is an input to MBROLA Text to Speech Synthesizer to produce "a" voice character during 51 ms, and to put a pitch of 114 Hz at 25% of said 51 ms.

(v) MBROLA Text to Speech Synthesizer (TTS)

MBROLA system is used to convert the phoneme command script into sound format (.wav) using Indonesian voice database that have been created by Arry Akhmad Arman. MBROLA is a text to speech synthesizer system (TTS) which are free and open source. In addition, the voice, that produced by MBROLA, should be intelligible and natural. There is a fundamental difference between TTS and any other talking machine (for example: cassette player). In the TTS it can automatically generates new sentences. Moreover, TTS isn't used for record all words of the focus language and that the biggest difference between TTS and particular talking machine such as voice response systems (Example: announcement arrival machine) which use a record of a word of focus language. Figure 3 shows a simple but general TTS system functional general diagram based on the work of Dutoit [6].

Natural Language Component converts digital text into phoneme code that contains a phoneme, pitch, and duration. Digital signal processing section will receive input of phoneme codes as well as pitch and duration generated by the previous section and based on these codes, the generator will produce speech sounds or speech signals corresponding to spoken sentences. Conversion from text to phonemes is influenced by the prevailing rules in a language (narrow phonetic transcription). In principle, this process is the conversion of textual symbols into phonetic symbols that represent the smallest unit of sound in a language. The way to read and pronunce text is very specific to each language. This causes the implementation of a text to phoneme converter unit becomes very specific to the language. To get more natural speech, speech that produced must have an intonation (prosody). Prosody is a conversion pitch value (basic frequency) during articulation process. After conversion in natural language component, digital signal processing component will receive input of phoneme codes as well as pitch and duration generated by the previous section and based on these codes, the generator will produce speech sounds or speech signals corresponding to spoken sentences.

III. EXPERIMENTAL RESULTS

The experiments conducted to evaluate the system by using two documents written in Bahasa Indonesia, which was printed out at resolution 300 dpi. The documents were "Pembukaan UUD 1945" and "Sumpah Pemuda".

In preprocessing phase, noises around the image edges caused by the scanner were eliminated. The process of elimination worked by checking edges of images based on the width and the height. If black pixels below threshold were found, they will be eliminated, vice versa. After line extraction completed, the pixels were be enhanced. In many cases the enhancement supported the segmentation result. It supported the segmentation, if the target contained blurred pixel which was often found in characters: "h","m","n". However, we did not obtain the benefit if the blurred pixels were those belong to noises. The result of character segmentation is shown at Tab.1

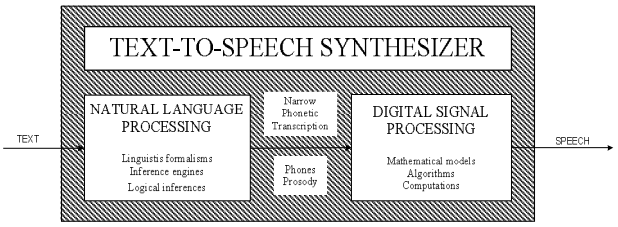

Figure 3: A simple but General TTS system functional diagram

TABLE 1 CHARACTER SEGMENTATION RESULT

	Pembukaan UUD	Sumpah Pemuda
Normal	1234 (100%)	221 (99%)
Num. of characters	1234 (100%)	222 (100%)
Punctuation marks	19	9
Words	178	35

"Pembukaaan UUD 1945" consists of 1234 characters and "Sumpah Pemuda" has 222 characters, including punctuation marks. The accuration result obtained by recognition module of "Pembukaan UUD 1945" was 78% (139 words correctly classified out of total 178), while "Sumpah Pemuda" was 57% (20 words correctly classified out of total 35). The errors were analyzed and we found that many of them were caused morphological similarity such as "e" was misclassified "o" character, "i" was misclassified as "l".

The errors were then corrected in the next phase: lexicon based post-processing using LCS algorithm and Eq.(1) was used to calculate the final score. Equation (1) makes the system to give higher score for the corrected word which has word length similar to the input word. Without Eq.(1), postprocessing applied to a word such as "adii" will obtain higher score if it is corrected as "di" (LCS score 1.0) compare to if it is corrected to "adil" (LCS score 0.75). Equation (1) weighted the original score by comparing the length of the word before and after correction, thus LCS score becomes 0.5 for "di" and 0.75 for "adil", thus "adil" is chosen and system produces correct choice.

After post-processing, the accuration result obtained for "Pembukaan UUD 1945" was 99%, while "Sumpah Pemuda" was 100%. The corrected words were inputted to TTS module which was implemented using MBROLA, resulting a voice in wav format.

IV. CONCLUSIONS

In this study, we developed "TextReader" system which combines Optical Character Recognition and Text To Speech Synthesizer. The system works by converting the image as a result of scanning to a textual documents into text (OCR phase), then followed by conversion from text to voice by TTS module. Using "TextReader", people with visual impairment will able to access textual information via voice, thus reducing their dependency to Braille characters.

"TextReaders" consists of preprocessing, segmentation, character classification, post-processing and text to speech synthesizer. The first four components are parts of OCR, while the last part is implemented using MBROLA TTS. The system was evaluated using two documents written in Bahasa Indonesia: "Pembukaan UUD 45" and "Sumpah Pemuda", yield an accuracy of 78% and 57%. Lexicon based preprocessing was implemented using Longest Common Subsequences algorithm which demonstrated significant improvement of the system accuracy by almost perfectly classified the characters. Text to Speech synthesizer converted the phoneme command script into voice in wav format.

Despiteof the success in this preliminary study, various improvements should be carried out to make the system ready to use by the visual impaired. Segmentation algorithm should be modified to handle multifont characters. Layout analysis should be improved to accurately locating the position of text and ignoring the non text region, either in grayscale or color background documents. Computational complexity of the lexicon based postprocessing should be reduced, by implementing more efficient algorithm compared to currently used linear search strategy. Inclusion of word frequency in the score calculation of postprocessing will be evaluated. Text to Speech Synthesizer should be improved so as the intonation of the output voice become natural. Interface of the future will be improved by including speech recognition module, thus enable the users to operate the system via voice, not through keyboard as implemented in this prototype.

ACKNOWLEDGEMENTS

The authors thank to Dr. Albertus Lukas for his advice and disccusion, and also to Dr. Arry Akhmad Arman who has permitted the authors to use diphone database for Bahasa Indonesia (id1) of MBROLA. This project is part of research activities in Media Understanding workpackage, under Ubiquitous e-Government Services Program (UGoS), Center for Information and Communication Technology, Agency for Assessment and Application of Technology, Indonesia (PTIK-BPPT).

REFERENCES

[1] A.S. Nugroho: Rehabilitasi Tuna Netra di Jepang: Survey penelitian dan kemungkinan aplikasinya di Indonesia , Proc. of Annual Seminar of Indonesian Student Society of Japan, Nagoya, Dec. 21 2002 (in Indonesian, available at http://asnugroho.net)

[2] K.Sawa, K.Magatani, K.Yanasima, "The Navigation System by Using Optical Beacon for The Visually Impaired", Proceedings of the 40th Conference The Japan Society of Medical Electronics & Biological Engineering, Nagoya, pp.293, May 2001

[3] One thousand books for visually impaired http://www.mitranetra.or.id/ebook/ last accessed (15 October 2010)

[4] A.S. Nugroho, S. Kuroyanagi, A. Iwata: An algorithm for locating characters in color image using stroke analysis neural network,Proc. of the 9th International Conference on Neural Information Processing (ICONIP'02), Vol.4, pp.2132-2136, November 18-22, 2002, Singapore

[5] N. C. Jones, P.A. Pevzner, Bioinformatics algorithms. Bradford 2004.

[6] T. Dutoit, An Introduction to Text-to Speech. Kluwer Academic, 1997