Intelligent Reading System Based on Mobile Platform

Kamil Aida-zade[1], Elshan Mustafaev[2], Jamaladdin Hasanov[3]
Cybernetics Institute of ANAS, Baku, Azerbaijan
[1]*kamil_aydazade@rambler.ru*, [2]*elshan_mustafayev@gmail.com*, [3]*jhasanov@gmail.com*

***Abstract*— This Article describes the work principles of the mobile platform based system of the printed text recognition and vocalization. The device consists of a camera-embedded mobile phone and special software for the recognition and speech synthesis of the Azerbaijani text. The offered device can be used for example, as a source of the information for visually impaired persons.**

Keywords— OCR; nueral networks; mobile; limited vision; disabled; speech synthesis; android

I. INTRODUCTION

One of the main problems for persons with limited vision is to independently get information from the traditional information sources like books, journals and magazines. Popular non-vision printed sources based on Braille's embossed dot system have certain disadvantages. Those books have bigger size due to embossed print and are easily damaged during the reading because of physical contact with paper. Another drawback is incomparably less quantity, compared to a common publishing. Additionally, the easiness of getting information is also cannot be compared to a common publishing.

Another popular way to get information is so called "audio book", where the texts of books, magazines, newspapers are stored in a musical compact disk or sound file. The advantages of this system are easiness to follow and fast reaction. Also in audiobooks the speed of the information exchange is sufficiently higher than reading. Despite these advantages, the main disadvantage of audio-books is quantity problem – only popular publishing and special information are issued in audio format.

As shown from above arguments, main requirements in the text to speech system are time (on demand access), no limit on information and the easiness of use. To fulfill the listed requirements, the camera embedded mobile computer might be required. This device should be independently run by user just by placing the reading material in front of the camera.

Currently, there are several software and device-based solutions for mobile reading. For instance, Intel has a product called Intel® Reader which consists of 5-megapixel camera with Intel processor and device for fixing of the reading material (Fig 1a). This camera has its own menu with easy navigation and can be used either independently or with adjusting device that can be assembled as a briefcase. The overall weight of the device is about 5.2 kg and duration of autonomous work is 4 hours. The price of the camera is 1,500 USD which is more expensive compared to personal computers.

Another similar reading solution is provided by Optelec and called ClearReader+(Fig 1b). This is a 2.5 kg portable device that consists of 5-megapixel camera and supports 12 languages.

Described systems have generally solved the problem but still have some drawbacks like weight, size, cost and of course problem of not supporting Azerbaijani language. This paper describes the solution that solves the mentioned problems with using the mobile phone as a recognition system platform.

a) b)

Figure 1. Text recognition systems. a) IntelReader; b) ClearReader+

The proposed software-device system consists of (1) a mechanical device to adjust the document for reading; (2) Android based mobile phone with camera and (3) software application for printed text recognition and speech synthesis. The system is designed for reading and playback of the Azerbaijani printed, hand-printed and handwritten texts. The described system operates in a following order: Reading material is fixed in a special place of the device and captured by the special camera. The captured image is sent to the input of the text recognition system. The result of the text recognition is transformed into a speech in a next layer, where it could be either played or saved.

II. SYSTEM OVERVIEW

Mobile recognition systems described in this article consist of following modules:

- Initial processing
- Segmentation

- Text recognition
- Lexicon analysis
- Speech synthesis

In initial processing module, the image taken from camera is processed and its orientation is corrected. Later, the image is converted into a black and white format and noises are removed. The corrected image is sent to the segmentation module where all text elements (text lines, columns, symbols, letters, etc.) are found and extracted. All text elements are indexed by their occurrence order in text (number of column, text line, word, etc.) and stored in a hash map for fast search and navigation. When pixels of the all symbols and their positions in text are known, the text recognition module is executed and image is converted into text.

The text recognition module consists of neural networks, each assigned for corresponding symbol type. All symbols analyzed for their baseline and sent to the input of the corresponding neural network. Neural network for each symbol type consists of 256 inputs which receive contour features from each image and activates the corresponding output indicating the character. The sequence of these characters are connected and compared in lexicon to either fix or accept the recognition result.

The result of the recognition module is analyzed by lexicon analysis module and the sequence of symbols assumed as a word is searched in lexicon database. Search method is based on Levenstein's "edit distance" which calculates the "distance" or "the count of symbols to be replaced" between two words. Ideally, in case of 100% recognition all words should match with database with zero character distance. If no word is matched, then the closest word from database is selected and suggested as a correct version of the recognized word. The lexicon database in basic model of the system consists of approximately 25,000 words that cover daily used words and their variations with suffixes and prefixes.

Final recognition results are grouped by sentences and sent to the input of the speech synthesis module. Speech synthesis module analyzes sentences and builds sound model based on the speech elements of the words. The final version is saved in WAV format and played when all processes are done.

III. Initial Processing And Text Recognition

The image captured by the camera phone will have different kinds of noises depending on camera angle, light, paper and text quality. Before the segmentation process, this image should be transformed into a black & white text with a proper orientation.

At the start of the image processing, its orientation should be found. The orientation of the paper is detected as edges of the white area which assumed as a background of the text. As mentioned above, the paper is not always associated with white color in images due to insufficient light or shadows. To correct this, the original image should be converted to black and white colors applying optimal thresholding method. In described system, the iterative method for automatic threshold detection described by Sonka, Hlavac and Boyle [1] and Parker [2] is used. The general algorithm of the iterative threshold determination method is described in fig 2.

Compute μ_1, the mean grey level of the corner pixels

Compute μ_2, the mean grey level of all pixels

$T_{old} = 0$

$T_{new} = (\mu_1 + \mu_2)/2$

while $T_{new} \neq T_{old}$ **do**

μ_1 = mean grey level of pixels for which $f(x,y) < T_{new}$

μ_2 = mean grey level of pixels for which $f(x,y) \geq T_{new}$

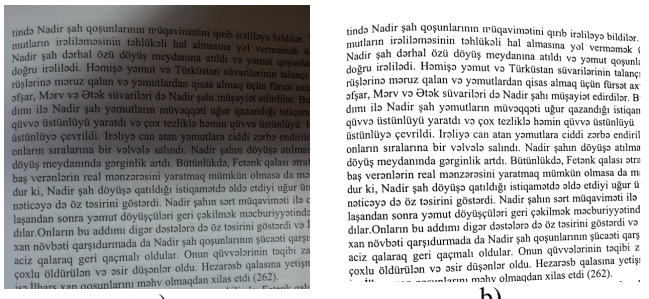
$T_{old} = T_{new}$

$T_{new} = (\mu_1 + \mu_2)/2$

end while

Figure 2. Algorithm of the iterative threshold detection method

The result of the iterative thresholding is shown in fig 3. Here, the initial image has a shadow on the right-top and its background is even darker than the fonts in right-bottom part (fig. 3a). In these cases, the iterative thresholding algorithm helps to convert the image into a black and white by processing it by small areas, instead of processing the whole area at once to find the threshold value in gray scale range of the image. As noticed, all regions of the image has been correctly converted into a black and white, regardless the excessive darkness in some areas (fig 3b).



a) b)

Figure 3. Result of the iterative threshold detection method.

In the next step, the search of the text elements in black and white image is started. The definition of the text elements is straightforward: text elements are union of black pixels that corresponds to a certain size. In other words, all printed elements, except lines, drawings, images, etc. are considered as text elements. All listed exceptions differ from text elements only by their size which in turn helps to avoid them easily considering only the element size. To filter elements by size, the threshold value of the size should be known. This threshold value calculation is based on the average value of the element height for each line. It might be calculated based on the whole document but many documents have different fonts in different

lines (like in text body and headlines) and calculating the average text element size for the all document can lead to undesired results.

The mentioned text element detection algorithm operates in a simple way: only 2 pixel lines of the image are scanned from top to bottom and iteratively next operations done:

1. The continuous black pixels on the first line are grouped and numbered with unique id (white pixels are assumed as delimiters).
2. Connections of black pixels in the second line to the grouped pixels in the previous (upper) line are detected and these pixels are marked with the same unique id.
3. Check for symbol "branching" – if two pixels having different unique id connects in second line (appearing in symbols like "u", "y", "x", "v", etc.), update the greatest id to a smallest one.

After the text element detection process, all pixel joints have their own identification id and by the position of the top, leftmost, rightmost and bottom pixels element borders are defined and the element size is calculated. In the next step, these elements are grouped by text line and their average size is calculated for this line as previously described.

When all image processing works are finished, each text element is associated with the next attributes:

1. Unique element identification number.
2. Number of column where element exists.
3. Number of line where element is located.
4. Number of word where element is located.
5. Index of symbol in word.
6. Size parameters.
7. Pixel data.

Listed attributes help to group and organize the result of the recognition module. Recognition module which consists of neural network module (NNM) receives the pixel data and returns the corresponding symbol in case of positive recognition.

The NNM consist of 4 neural networks which are dedicated for certain type of symbols. All symbols are divided into next types, according to baseline detection rules [3]:

1. Capital letters and digits.
2. Symbols having ascending strokes (like "d", "t" and "b").
3. Symbols having descending strokes (like "q", "y" and "p").
4. Symbols without ascending and descending strokes (like "a", "o" and "e").

The reason to split the neural networks into 4 parts is to increase the precision and speed of the recognition. In case of using 1 neural network, it should has 76 (64 letters (32 capital + 32 small), digits from 0 to 9 and special signs) neurons in output which will take more time to train and recognize than neural networks with 6 (symbols with ascending strokes) or 24 outputs (symbols with no strokes).

Another benefit of separately recognizing symbols is isolating similar looking symbols like "l" (12th letter of English alphabet), "1" (one) and "I" (capital "i").

TABLE I. SAMPLE OF SAME LETTERS HAVING DIFFERENT FORM

Arial	Times New Roman	Courier New
f t g l	f t g l	ftgl

Additionally, different NNMs per font were created. It helps to increase the quality of recognition which can be reduced by representing the same object with different samples. In table 1, representation of same letters in Arial, Times New Roman and Courier New are shown. As shown from the table, symbols have different strokes which make them look different.

The neural network used in this system is a 3-layered perceptron that has 256 neurons in its input. The inputs are feed with 256 PDC features taken from the normalized to 32x32 pixel image of the text element [4,5].

The results of the text recognition are matched with lexicon database and the final result is sent to the input of the speech synthesis system.

IV. SPEECH SYNTHESIS

Speech synthesis block of the system interprets the text information into a human speech. Initially, all words are split into vowels and then concatenated by synthesis rules.

The rough, primary basis of a formed acoustic signal is created on a basis of concatenation of the fragments of an acoustic signal taken from speech of the announcer – a "donor". Further this acoustic basis is exposed to updating by the rules, function of which consists of giving the necessary prosodies characteristics (frequency of the basic tone, duration and energy) to the "stuck together" fragments of an acoustic signal.

In systems of concatenate synthesis (earlier it was called compilation), synthesis is carried out by fuse necessary units from available acoustic stock. Concatenate of segments of the written down speech lays in a basis of concatenate synthesis. As a rule, concatenate synthesis gives naturalness to sounding of the synthesized speech. Nevertheless, the natural fluctuations in speeches and the automated technologies of segmentation of speech signals create noise in the received fragment, which reduce naturalness of sounding.

Formant synthesis does not use any samples of human speech. On the contrary, the speech message of the synthesized speech is created by means of acoustic model. Parameters, as own frequency, sounding and noise levels vary after the lapse of time and are created the form of a signal of artificial speech.

The method of concatenation at an adequate set of base elements of compilation provides qualitative reproduction of spectral characteristics of a speech signal, and the set corrected-possibility of formation natural intonation-prosodial registrations of statements.

An acoustic signal database (ASD), which consists of fragments of a real acoustic signal - elements of concatenation (EC) is the basis of any system of synthesis of the speech based on concatenation a method. Dimension of these elements can be various depending on a concrete way of synthesis of speech, it can be phonemes, allophones, syllables, diaphones, words and et cetera [6].

Generated speech signals are saved as a multimedia file format and played using Android's multimedia API.

V. RESULTS

Described recognition system has been tested for functionality with Azerbaijani texts written in A4 paper with 12, 14 and 16 sized Arial, Times New Roman and Courier New fonts. The result of the recognition and speech synthesis of the texts covering different areas and topics, including symbols, digits and signs was successful. The quality of the recognition and synthesis was as described in works [4,5,6,7].

REFERENCES

[1] Milan Sonka, Vaclav Hlavac, and Roger Boyle. Image Processing, Analysis and Machine Vision. International Thomson Computer Press, 1993.

[2] J. R. Parker. Algorithms for Image Processing and Computer Vision. John Wiley & Sons, 1997.

[3] K.R. Aida-zade, J.Z. Hasanov. "On initial processing method in handwritten text recognition systems". "Problem of Cybernetics and Informatics", Volume III. 24-26 October, 2006. Baku, Azerbaijan. PCI2006.

[4] K.R. Aida-zade, E.E. Mustafayev. "Clusterization of the hand-printed azerbaijani alphabet using self-organizing Cohonen map". Proceedings of the scientific conference "Modern problems of informatics, cybernetics and information technologies". Volume I, Baku, Azerbaijan, 2004.

[5] K.R. Aida-zade, E.E. Mustafayev. "On parameter optimizaion of neural networks in training stage". Proceedings of the scientific conference "Modern problems of informatics, cybernetics and information technologies". Volume I, Baku, Azerbaijan, 2003.

[6] K.R. Aida-Zade, A.M. Sharifova. "Analysis of approaches to text to speech synthesis and their application to Azerbaijani language". The second international conference "Problem of Cybernetics and Informatics" dedicated to the 50th anniversary of the ICT in Azerbaijan, September 10-12, 2008. Baku, Azerbaijan PCI2008.