ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 6, December 2014

Design and Implementation of Speech Generation System using MATLAB

Pooja Chandran, Aravind S, Jisha Gopinath and Saranya S S
Department of Electronics and Communication Engineering, Sree Buddha College of Engineering for Women, Kerala, India

Abstract- The main idea of this project is to recognize the text character and convert it into speech signal. The text contained in the page is first pre-processed. The pre-processing module prepares the text for recognition. Then the text is segmented to separate the character from each other. Segmentation is followed by extraction of letters and resizing them and stores them in the text file. These processes are done with the help of MATLAB. This text is then converted into speech.

***Index terms*- OCR, Segmentation, Templates, TTS.**

I. INTRODUCTION

Speech is probably the most efficient medium for communication between humans. Speech synthesis is the artificial synthesis of human speech [1]. A text-to-speech synthesizer (TTS) is a computer based system that should be able to read any text aloud, whether it is directly introduced in to the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. A TTS converts normal language text into speech whereas other systems render symbolic linguistic representations like phonetic transcriptions into speech. OCR system can be used as an input device for reading a text [2]. Different recognition engines include OCR, OMR, ICR, MICR, Scanners and Camera etc. Here we are using scanner for reading text or a low cost camera can also be used. Character recognition or optical character recognition system converts scanned text or images into a computer format text. The three main elements for text reading technology include: scanning, recognition and reading text. Text to speech conversion systems have an enormous range of applications. Their first real use was in reading systems for the blind, where a system would read text from a book and convert it into speech.

II. HISTORY OF TTS SYSTEM

Long before electronic signal processing was invented, there were those who tried to build machines to create human speech. Some early legends of the existence of speaking heads involved Gerbert of Aurillac (d.1003 AD), Albertus Magnus (1198-1280), and Roger Bacon (1214-1294). The first computer based speech synthesis systems were created in the late 1950s, and the first complete text-to-speech system was completed in 1968. The first successful prototype of reading machine was developed at Haskin Laboratories in 1970s. These large prototypes sent the output from a fixed font OCR to the input of speech synthesis algorithm developed at Haskin Laboratories. Available text reading systems for visually impaired persons are Cicero text reader, i-scan, Kurzweil 1000, Open Book, Ovation, Scan N Talk, VERA etc. All these systems are highly expensive and cost more than 2 lakh INR [3]. By the end of 2009, Intel announced a Linux based text reading device with optical character recognition and text-to-speech technology. But the cost of the system is nearly 80000 INR. Another text reader has also been introduced in 2010, Optelec Clear Reader+ Advanced, a reading assistant to scan, view, magnify, save and listen to any printed text in one portable solution [4]. The cost of the system is above 1lakh INR. Early electronic speech synthesizers sounded robotic and were often barley intelligible. The quality of synthesized speech has steadily improved, but output from contemporary speech synthesis systems is still clearly distinguishable from actual human speech. Existing systems for text recognition are typically limited either by explicitly relying on specific shapes or colour masks or by requiring user assistance or may be of high cost. Therefore we need a low cost system that will be able to automatically locate and read the text aloud to visually impaired persons.

III. OVERVIEW OF TEXT PROCESSING

The first step in the process is to digitize the analog document using a digital scanner, and then extracted text will be pre-processed. When the regions containing text are located, each symbol is extracted through a segmentation process.

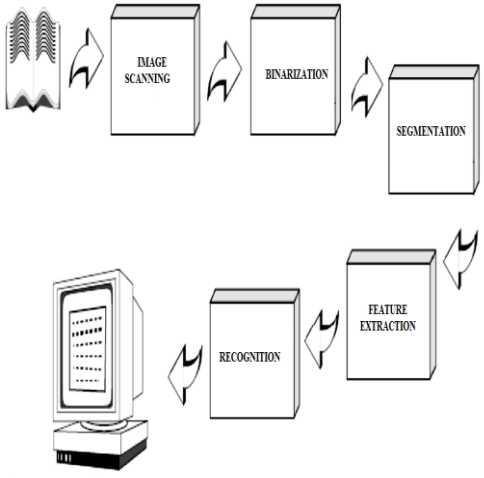

Fig 1: Components Of Text Processing System

The identity of each symbol is found by comparing the extracted features with the descriptions of the symbol classes obtained through a previous learning phase [5]. Finally contextual information is used to reconstruct the

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 6, December 2014

words and numbers of the original text. The recognition of scanned document images using Optical Character Recognition (OCR) is now generally considered to be a solved problem for some scripts. OCR converts images of documents that are printed or hand written into recognized characters. Components of an OCR system consist of optical scanning, binarization, segmentation, feature extraction and recognition.

A. Image Scanning

In computing a scanner is a device that optically scans images, printed text, handwriting, or an object, and converts it to a digital image. Different types of scanners are hand-held scanners, mechanically driven scanners, rotary scanners, planetary scanners and digital camera scanners.

B. Binarization

Binarization is the process of converting a gray scale image into binary image by thresholding. The binary document image allows the use of first binary arithmetic during processing, and also requires less space to store. Because of the complexity of the OCR operation, the input of the character recognition phase in most methods is binary images [6]. Therefore, in the pre-processing phase, gray scale images are to be converted to binary images.

C. Segmentation

Segmentation of text is a process by which the text is partitioned into its coherent parts [7]. The text image contains a number of text lines. Each line again contains a number of words. Each word may contain a number of characters. The segmentation schemes such as line segmentation, word segmentation and character segmentation are proposed where lines are segmented then words and finally characters. These are then put together to the effect of the recognition of individual characters.

D. Feature Extraction

The objective of feature extraction is to capture the essential characteristics of the symbols, and it is generally accepted that this is one of the most difficult problems of pattern recognition [8]. The most straight forward way of describing a character is by the actual raster image. Another approach is to extract certain features that still characterize symbols, but leaves out the unimportant attributes. The five main techniques for extraction of features are:

1) Template matching and correlation techniques.
2) Feature based techniques.
3) Distribution of points.
4) Transformations and series expansions.
5) Structural analysis.

E. Recognition

After we got the character by character segmentation we store the character image in a structure. This character has to be compared with the pre-defined character set. Preliminary data will be stored for all characters for an identified font and size. This data contains the following information: character ASCII value, character name, and character BMP image, character width etc. for every recognized character above mentioned information will be capture. The recognized character information will be compared with the pre-defined data which we have stored in the system. As we are using the same font and size for the recognition there will be exact one unique match for the character. This will identify us the name of the character [7]. If the size of the character varies it will be scaled to the known standard and then recognizing process will be done.

IV. SPEECH SYNTHESIS

Speech synthesis is the artificial production of human speech. A system used for this purpose is called a speech synthesiser and can be implemented in software and hardware. Synthesised speech can be created by concatenating pieces of recorded speech that are stored in a database [1]. Systems differ in the size of stored speech units, a system that stores phones or diaphones provides the largest output range, but may lack clarity. For specific usage domains the storage of entire words or sentences allows for high quality outputs. Alternatively a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely synthetic voice output. Many computer operating systems have included speech synthesizers since the early 1980s. The quality of a speech synthesizer is judged by its similarity to the human voice, and by its ability to be understood. An intelligible text to speech program allows people with visual impairments or reading disabilities to listen to written words works on a home computer.

V. TEXT TO SPEECH SYNTHESIS

A Text to speech synthesizer is used to define Text-To-Speech as the automatic production of speech, through a grapheme-to-phoneme transcription of the sentences to utter. The functional diagram is shown in Fig 2.

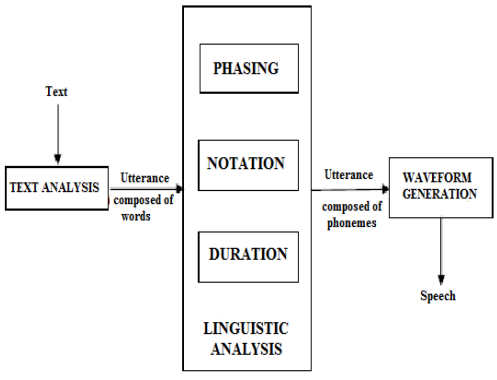

Fig 2: Functional block diagram

The functional diagram of a very general TTS synthesizer comprises of Natural Language Processing module (NLP) and a Digital Signal Processing module (DSP).NLP is capable of producing a phonetic transcription of the text read, together with the desired intonation and rhythm (often termed as prosody).DSP module which

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 6, December 2014

transforms the symbolic information it receives into speech [9].

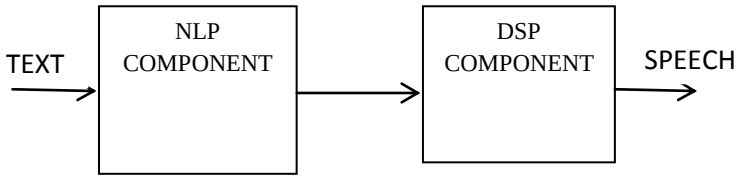

Fig 3: TTS Synthesizer

1) The NLP Component

The NLP module takes a series of text input and produces a phonetic transcription together with the desired intonation and prosody (rhythm) that is ready to pass on the DSP module. The NLP module is composed of three major components: text analyser, letter-to-sound (LTS), and prosody generator. Text analysis is a language-dependent component in TTS system. It is invoked to analyse the input text. This process can be divided into three major steps: Pre-processing, Morphological analysis and Contextual analysis. During Pre-processing stage, the main components of the input text are identified [6]. Usually, pre-processing also segments the whole body of text into paragraphs and organizes these paragraphs into sentences. Finally, pre-processing divides the sentences into words. The morphological analysis serves the purpose of generating pronunciations and syntactic information for every word in the text. Morphological analysis determines the root form of every word and allows the dictionary to store just headword entries, rather than all derived forms of a word. Contextual analysis considers words in their context and determines the part-of-speech (POS) for each word in the sentence. Context analysis is essential to solve problems like homographs (words that are spelled the same way but have different pronunciations).Letter-To-Sound (LTS) module is responsible for automatically determining the incoming text's phonetic transcription. Prosodic or supra-segmental features consist of pitch, duration, and stress over the time. Prosodic features can be divided into several levels such as syllable, word, or phrase level [10].

2) Digital Signal Processing (DSP) Module

The operations involved in the DSP module are the computer analogue of dynamically controlling the articulatory muscles and the vibratory frequency of the vocal folds so that the output signal matches the input requirements. This can be basically achieved in two ways: Explicitly, in the form of a series of rules which formally describe the influence of phonemes on one another and Implicitly, by storing examples of phonetic transitions and co-articulations into a speech segment database, and using them just as they are, as ultimate acoustic units [10].

VI. SYSTEM IMPLEMENTATION

The proposed system is implemented using MATLAB.

Operational stages of the system can be summarized as:

1. Image capture
2. Image pre-processing
3. Image filtering
4. Character recognition
5. Text to speech conversion.

Step 1:- First, image captured by camera is stored in a file in a specified format. Then the image is read with the help of imread command. imread Read image from graphics file.

Step 2:- Second step is pre-processing step. In this step firstly we convert the image into gray scale by rgb2gray command. rgb2gray convert RGB image or colormap tograyscale. rgb2gray converts RGB images to grayscaleby eliminating the hue and saturation information while retaining the luminance.fid = fopen (filename) opens the file filename for read access. Filename is a string containing the name of the file to be opened. Use an empty matrix to store the recognized characters.

Step 3:- In this step we perform image filtering. Image filtering comprises of cropping the image so that text only areas are retained. This can be done by finding row and column indices and cropping out the regions outside maximum value of row and column indices.

Steps 4-10:- Next is character recognition. Character recognition comprises of cropping each line, cropping each letter, performing correlation and writing to a text file. Before performing correlation we have to load the already stored model templates so that we can match the letters with the templates.

Step 11:- Now we have to convert this text to speech. For this, first we analyze the text, and check the condition that if Win 32 SAPI is available in the computer or not. If it is not available then error will be generated and we should load that Win 32 SAPI library in the computer.

Step 12:- This step will be executed if there is Win 32 SAPI file in the computer. Then in this step we make a new server for this file by actxserver command.

Step 13:- In this step we get voice object from Win 32 SAPI by invoke command

Step 14:- In this step we compare the input string with Win 32 SAPI string with strcmp command.

Step 15:- In this step we extract voice by firstly select the voice which are available in that library

Step 16:- In this step we choose the pace of voice, by using the commands, nargin and nargout.

Step 17:- In this step we initialize the wave player for convert the text into speech by convert uint8 to double precision.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 6, December 2014

Step 18:- Finally we get the speech for given image.

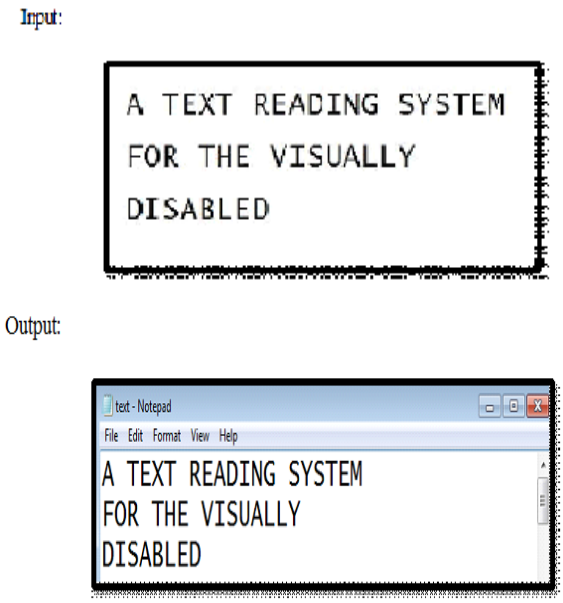

Fig 4: Input and Output of the implemented system

VII. APPLICATIONS

1) Aid to handicapped persons

Voice handicaps originate in mental or motor/sensation disorders. With the help of an especially designed keyboard and a fast sentence assembling program, synthetic speech can be produced in a few seconds to remedy these impediments [1].

2) Talking books and toys

The toy market has already been touched by speech synthesis. Many speaking toys have appeared, under the impulse of the innovative 'Magic Spell' from Texas Instruments [1].

3) Vocal Monitoring

In some cases, oral information is more efficient than written messages. The appeal is stronger, while the attention may still focus on other visual sources of information. Hence, the idea of incorporating speech synthesizers in measurement or control systems.

4) Multimedia, man-machine communication

In the long run, the development of high quality TTS systems is a necessary step towards more complete means of communication between men and computers. Multimedia is a first but promising move in this direction [1].

5) Fundamental and applied research

TTS synthesizers possess a very peculiar feature which makes them wonderful laboratory tools for linguists: they are completely under control, so that repeated experiences provide.

6) Language education

High Quality TTS synthesis can be coupled with a Computer Aided Learning system, and provide a helpful tool to learn a new language.

VIII. CONCLUSION

This paper is an effort to suggest an approach for image to speech conversion using optical character recognition and text to speech technology. The application developed is user friendly, cost effective and applicable in the real time. By this approach we can read text from a document, Web page or e-Book and can generate synthesized speech through a computer's speakers or phone's speaker. The developed software has set all policies of the singles corresponding to each and every alphabet, its pronunciation methodology, the way it is used in grammar and dictionary. This can save time by allowing the user to listen background materials while performing other tasks. System can also be used to make information browsing for people who do not have the ability to read or write. This approach can be used in part as well. If we want only to text conversion then it is possible and if we want only text to speech conversion then it is also possible easily. People with poor vision or visual dyslexia or totally blindness can use this approach for reading the documents and books. People with speech loss or totally dumb person can utilize this approach to turn typed words into vocalization. Experiments have been performed to test the text reading system and good results have been achieved.

REFERENCES

[1] N.Swetha, K.Anuradha,"Text To Speech Conversion", IJATCSE, Vol.2, No.6, 2013.

[2] Smith R., "An overview of the Tesseract OCR Engine, ICDAR 2007. Vol.2 ,2007

[3] http://www.linuxdevices.com.

[4] http://www.optelec.com.

[5] Jianli Liu.Nugent, J.H.Bowen, "Intelligent OCR editor", Canadian Conference, vol.1, 1993.

[6] D.Sasirekha, E.Chandra,"Text To Speech: A Simple Tutorial", IJSCE, Vol.2, Issue-1, 2012.

[7] Aparna.A, I.Muthumani," Optical Character Recognition for Handwritten Cursive English characters", IJCSIT, Vol.5 (1), 2014.

[8] Sukhpreet Singh, "Optical Character Recognition Techniques: A Survey", IS Journal, Vol. 4, No.6, 2013.

[9] Dutoit, Thierry. "An Introduction to Text-To-Speech Synthesis," Boston: Kluwer Academic Publishers, 1997.

[10] Y. Sagisaga, "Speech Synthesis from Text," 1998.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 6, December 2014

AUTHOR BIOGRAPHY

Aravind S, Assistant professor in the Department of Electronics and Communication Engineering,Sree Buddha College of Engineering for Women, Mahatma Gandhi University, Kerala, India. He obtained M.Tech degree in VLSI and Embedded Systems from College of Engineering Chengannur, Cochin University in 2012. He received his B.Tech Degree in Electronics and Communication Engineering from Cochin University, Kerala, India, in 2009. His area of interest include Network Theory, Signals and Systems, Embedded Systems, Digital Electronics, DSP and VLSI.

Pooja Chandran, pursuing final year BTech degree in Electronics and Communication Engineering from Mahatma Gandhi University, Kerala, India.

Jisha Gopinath, pursuing final year BTech degree in Electronics and Communication Engineering from Mahatma Gandhi University, Kerala, India. Completed Diploma in Electronics Engineering from Technical Board of Education, Kerala

Saranya.S.S, pursuing final year BTech degree in Electronics and Communication Engineering from Mahatma Gandhi University, Kerala, India. Completed Diploma in Electronics Engineering from Technical Board of Education, Kerala.