Available at: www.researchpublications.org

Text to Speech Converter

A marvel for visually challenged people

Sagar S. Alande[1],Soumi S. Sharma[2], Ajita A. Chavan[3]

[1&2&3] Electronics Department, KJSIEIT, Sion, Mumbai, India
E-mail: sagar.alande@somaiya.edu [1],soumi.sharma@somaiya.edu [2], ajita.chavan@somaiya.edu[3]

***Abstract* –A text to speech system (TTS) converts normal language text into speech. An intelligible system allows people with visual impairments or reading disabilities to listen to written works. The proposed system has been the hardware solution for synthesizing speech thus enabling access to digital content in voice mode. Our project has been categorized into two, image processing and speech processing. The main objective is to provide operands recognised as alphabets and numbers using a camera, followed by segmentation and feature extraction in Matlab. The text is sent to the DSP processor *TMS320C6713* which synthesizes the data, disintegrates into allophones and provides a voice output of the input text after image processing.**

***Keywords* – TTS, Segmentation, Template Matching, Speech Synthesis, DSP *TMS320C6713* Kit**

I. INTRODUCTION

Text to speech is a process through which text is rendered as digital audio and then spoken. A text to speech synthesizer is a computer based system that can read text aloud automatically, regardless of whether the text is introduced by a computer input stream or a scanned input submitted. A speech synthesizer can be implemented by both hardware and software. A rapid improvement has been made in this field over the couple of decades and lot of high quality TTS systems are available for commercial use. [14]

As there are number of research prototypes of TTS [5] systems developed and none was compared with the commercial grade TTS systems for quality. The main reason is that it needs improvisation in collaboration between linguistics and technologists. Text to speech should be efficient to communicate information to the user, when digital audio recordings are inadequate, for developing a user friendly speech synthesizer. In this way the system widely helps in developing a Computer-Human interaction like-voice annotations to files, Speech enabled applications, talking computer systems (GPS, Phone-based) etc. [16]
Our Text to Speech converter accepts a string of 36 characters of text (alphabets and/or numbers) as input. In this we have taken the images of printed content of text and performed various steps of image processing. The objective is to recognize the text and graphics contents in the images and extract the intended information as human would. Character extraction is the extraction of characters from document images and analysis of the same. It is one of the key tasks of document image analysis involving denoising, segmentation, feature extraction and template matching. [15]

Rest of the paper is organised as follows: section II deals with the architecture of the TTS system, section III deals with the principle of the steps in the TTS system, section IV describes the methodology followed by algorithm in section V, future prospects are discussed in section VI and finally conclusion of the paper is presented in section VII.

II. ARCHITECTURE OF TTS

The TTS system consists of 5 fundamental components [13]
A. Text Analysis and Extraction
B. Text Detection
C. Phonetic Analysis
D. Prosodic Modelling and Intonation
E. Acoustic Processing

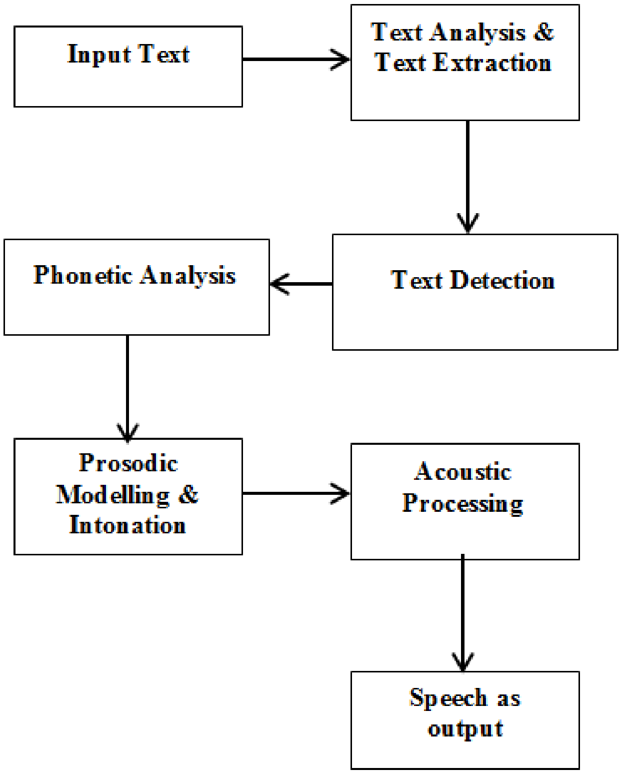


Fig.1 System overview of TTS

Thus in this way the input text is passed through these phases to obtain the speech as an output.

III. PRINCIPLE

A. Text Analysis and Extraction

The Text Analysis part is pre-processing part which analyse the input text. The input image can be colour or gray scale image [4]. The proposed filter decomposes input image into several sub images based on the size of the characters. Extraction processing includes three steps – feature emphasis, character extraction and noise reduction. In feature emphasis [11] step, among various filters like Gaussian, Laplacian of Gaussian filters, Gaussian filter is used to emphasize character features in sub images and removes most of the noise.

Text extraction [6] is done in order to speed up the data entry. This paper aims at performing filtering of the taken images through camera followed by text extraction and localization through segmentation and other morphological techniques. This paper describes the edge based technique for text detection. Edge based methods focus on high contrast between the text and the background and the edges of the text boundary are identified and merged. [1], [2]

B. Text Detection

Text detection is done under the heading template matching including the formation of the template database [3]. Matching is based on polygon matching, formation of chain codes or segmenting each character based on the number of pixels assigned for each alphabet.

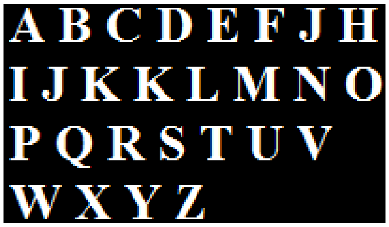


Fig.2 Samples of Template

C. Phonetic Analysis

Phonetic Analysis converts the orthographical symbols into phonological ones using a phonetic alphabet, basically known as “grapheme-to-phoneme” conversion. Phone is a sound that has definite shape as a sound wave. Phone is the smallest sound unit. A collection of phones that constitute minimal distinctive phonetic units are called Phoneme. [17]

D. Prosodic Modelling and Intonation

The concept of prosody is the combination of stress pattern, rhythm and intonation in a speech. The prosodic modelling describes the speaker’s emotion. Intonation is simply a variation of speech while speaking.

E. Acoustic Processing

The speech will be spoken according to the voice characteristics of a person, there are three type of Acoustic synthesing available.

A. Concatenative Synthesis

The concatenation of pre-recorded human voice is called Concatenative synthesis. In this process a database is needed having all the pre-recorded words.

B. Formant Synthesis

Formant-synthesized speech can be constantly intelligible. It does not have any database of speech samples.

C. Articulatory Synthesis

Speech organs are called Articulators. In this articulatory synthesis techniques for synthesizing speech based on models of the human vocal tract are to be developed. It produces a complete synthetic output. [13]

IV. METHODOLOGY

In this paper the camera-captured document image is considered as the input. The document image may consist of alphabets and numerals. First step is to read the input image that is the scanned or camera captured. The captured image is loaded as the input. If coloured then the image is converted from RGB to binary in the pre-processing stage called as binarization and then the background is removed. The image resulting from the scanning process may contain a certain amount of noise and hence, pre-processing is required. Gaussian filter is used for the same.

The relevant information about the characters is extracted from the pre-processed image using erosion and dilation related morphological techniques and segmentation is done. The images of A to Z and numerals 0 to 9 are set as templates. They are resized to the dimensions of the extracted characters and template matching is done. The matching is done on the basis of boundary detection. Thus, polygon detection is used. The characters in the template are assigned particular polygons and then they are matched accordingly.

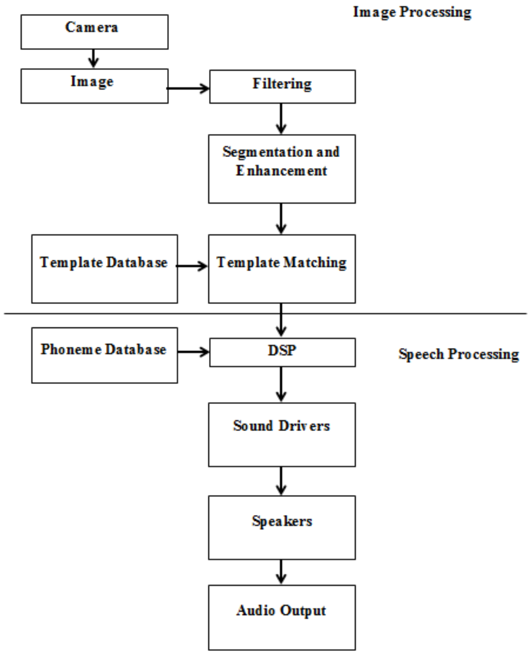


Fig.3 Methodology

DSP *TMS320C6713DSK* [10] is used for the speech [8] processing part. Phoneme database is created by recording of every alphabets and numerals called phones and converting it into a wave file [9]. The speech waveform is converted to a type of parametric representation for further analysis and processing. This is referred to as the signal-processing front end ** followed by the matching of the template. The matched waveform is identified and the particular file is read aloud through the DSP. Sound drivers are also used to amplify the sound and the audio is obtained as output.

V. ALGORITHM

A. Image Processing

In this paper, the captured image is filtered initially before processing the image to remove the noise present in the image. 3*3 Gaussian filter is used for this purpose.
Gaussian will have function as follows:

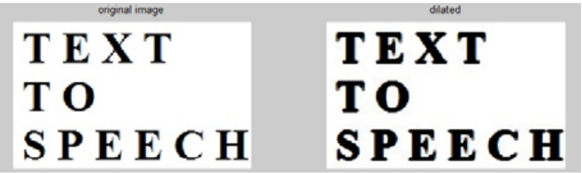
$$H\ (m, n) = e^{-\left\{\frac{(m+n)^2}{2\sigma^2}\right\}}$$

3*3 mask is given by: $H(m,n) = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$

The filtered image is dilated and eroded for further enhancement. Dilation [3] is a process in which the binary image is expanded from its original shape using a structuring element. The dilation operation is defined by

$$X \oplus A = \{z|\ [(\hat{A})_z \cap X] \subseteq X\}$$

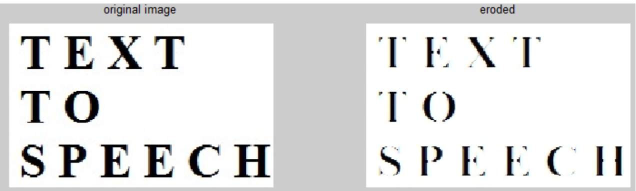
Where $(\hat{A})_z$ = the image A rotated about origin.



Erosion [3] is a counter process of dilation. If dilation enlarges image then erosion shrinks the image. The way image is shrunk is determined by the structuring element. The erosion operation is defined by

$$X \square A = \{z|\ (A)_z \subseteq X\}$$

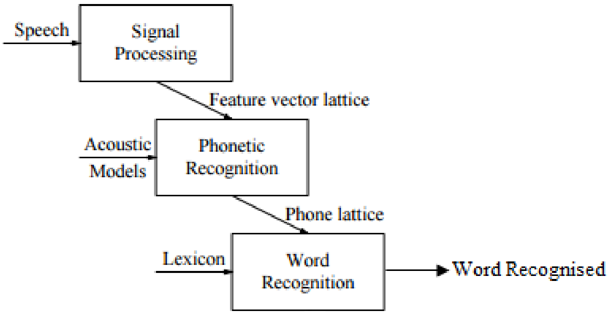
Where $(A)_z$ = the image A rotated about origin



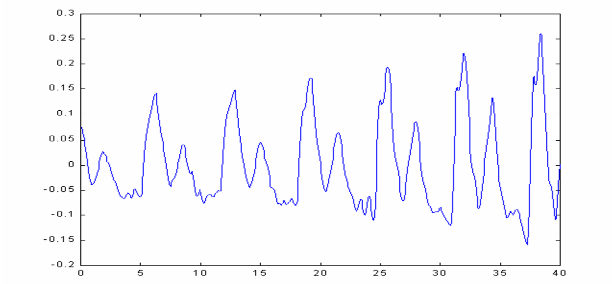
Representation techniques are used for the templates so that matching can be made more reliable. The representation technique used in this paper is Polygon Approximation. Polygons for various characters are set and then compared with the extracted character. Maximum matched character is finally identified.

B. Speech Processing

In this paper the processed image is matched with the phoneme database. Phoneme database is created by recording each of the characters and then converting it to wave file. The MFCCs [18] are calculated from the waveforms which are used for matching. Matching is done by finding out the Euclidean distance between the input signals and the voice signals in the database. Finally, audio output is obtained.



Linear Prediction Coding (LPC) methods are used in speech coding, speech synthesis, speech recognition, speaker recognition and verification and for speech storage. They provide extremely accurate estimates of speech parameters.



VI. FUTURE WORK

Direct reading of text from binary image or colour image is a challenging task [7] because of the complex background or degradations introduced during the scanning of a paper document. In this paper a simple solution is presented based on dilation and erosion. In the future, the proposed work can be enhanced by increasing the database to contain alphabets, numbers of any font style, font size and making friendlier by including different handwritten styles. However the one disadvantage of having both numbers and alphabets in the database is the possibility of misinterpretation of digits having similar features. This can be overcome by improving the morphological techniques used.
Even more variation under the part of prosodic modelling and intonation can be done in the TTS system. Prosody is the combination of stress pattern, rhythm and intonation in a speech. Thus by introducing prosodic modelling speaker's

emotions can also be given equal priority and in turn use of TTS systems will increase drastically.

Finally, the system can be enhanced to be able to read a book as efficiently as it is doing with few alphabets and numerals. For reading a book, equipment must be made in such a way that it is able to turn the page and go to the next page and start reading the text.

VII. CONCLUSION

In this paper, a simpler and easy to implement technique is used to extract and localize the captured image through MATLAB. The method adopted can read the characters from camera captured efficiently with good accuracy. The TTS [5] system prepared also serves to be portable and handy for many. This system also proves to be cost-effective. Finally, all related methods given in references are analysed and the drawbacks are reduced [3] and thereby getting an improved version of the previous works.

REFERENCES

[1] Giri, P. S., *"Text Information Extraction and analysis from Image Using Digital Image Processing Techniques"*, 2nd edition [2013].

[2] M. Swamy Das, B. HimaBindhu, A. Govardhan, *"Evaluation of Text Detection and Localization Methods in Natural Images*", 3rd edition[2012].

[3] Rafael Gonzales & Richard.E.Woods,*"Digital Image Processing*", Pearson education, 2nd edition 2001.

[4] Agarwal P, Varma R *"Text Extraction from Images",* IJCSET[2012].

[5] Tatham, M.; Lewis, E., *"Improving text-to-speech synthesis",* Spoken Language. ICSLP 96. Proceedings, Fourth International Conference on, vol.3, no., pp.1856-1859 vol.3, 3-6 Oct 1996.

[6] Park, C. J., Moon, K. A., Oh, Weon- Geun, and Choi, H. M.. An efficient of character string positions using morphological operator. IEEE International Conference on Systems, Man, and Cybernetics, 3, 8-11: 1616-1620[2000].

[7] Jagath Samarabandu,*"An Edge-based Text Region Extraction Algorithm for Indoor Mobile Robot Navigation"* International Journal of Signal Processing 3(4)2007.

[8] *URL http://www.w3.org/TR/jsml/*

[9] Speech recognition using DSP, www.vgyan.com/seminar/download/

[10] *"TMS320C6713 Floating point Digital Signal Processing data sheet,"* Texas Instruments, Houston.

[11] Silvio Ferreira, Celina Thillou, Bernaud Gosselin, *"From Picture to Speech: an Innovative Application for Embedded Environment".*

[12] Jindrich Matousek, Josef Psutks, Jiri Krita, *"Design of speech Corpus for Text-to-Speech Synthesis*".

[13] Text To Speech: A Simple Tutorial, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[14] Rehna. V. J, R. Neha & Sampada. H. K, *"Character extraction and recognition from document images using segmentation and feature extraction"*, IRNet Transactions on Electrical and Electronics Engineering (ITEEE) ISSN 2319 – 2577, Vol-1, Iss-2, 2012.

[15] Rishpa Sachdeva, Puja Nagpal, "*Text Localization and Extraction in Images Using Mathematical Morphology and OCR Techniques*", International Journal of Scientific Engineering and Research (IJSER), Volume 1 Issue 1, September 2013.

[16] Subbarao.Y.K, J.S.Chitode, *"Text-To-Speech Implementation in Field Programmable Gate"* Array, National Conference On Advanced Computing And Computer Networks (NCACCN) 2007.

[17] Milind U. Nemade*, Satish K.Shah, *" Real Time Speech Recognition Using DSK TMS320C6713", ISSN: 2277 128X,* International Journal of Advanced Research in Computer Science and Software Engineering, *Volume 4, Issue 1*, January 2014,

[18] Sneha Hegde, Amruta Pendharkar, Prathamesh Pewekar, Aniruddha Satoskar, *"Speaker Recognition Using TMS320C6713DSK*", Department of Electronics Engineering, Sardar Patel Institute of Technology, University of Mumbai, 2008-2009.

[19] *http://www.dsptutor.freeuk.com/analyser/guidance.html#leakage.*

[20] *TMS320C6713DSK*, Technical Reference, 2003, Printed in 2003