REVIEW OF TEXT TO SPEECH CONVERSION METHODS

[1]POONAM.S.SHETAKE, [2]S.A.PATIL, [3]P. M JADHAV

[1,3]Department of Electronics, TEI, Rajwada Ichalkaranji
[2]Department of E & TC, TEI, Rajwada Ichalkaranji
E-mail: [1]poonamshetake3@gmail.com, [2]shrinivasapatil@gmail.com, [3]pmjadhav85@gmail.com

Abstract- A text to speech converter convert's normal language text into speech. Text to speech converter is useful in different applications. Customer support dialog systems Interactive voice response (IVR) systems etc and are also useful in an applied research. This application is more helpful in banking, toys and many other applications like checking marks, railways, aid to the physically challenged persons, language education and fundamental and applied research. etc. But text to speech conversion is not that much easy for machine as it is for human. Basic steps that machine has to follow for text to speech analysis are database creation, character recognition and text to speech conversion. This paper surveys methods related to character recognition as well as approaches used for text to speech conversion for machine.

Keywords- OCR, Text to Speech phoneme diphone

I. INTRODUCTION

Language is the ability to express one's thoughts by means of a set of signs, whether graphical gestural, acoustic, or even musical. It is distinctive nature of human beings, who are the only creatures to use such a structured system. Speech is one of its main components. It is by far the oldest means of communication between human being and also the most widely used. No wonder, then, that people have extensively studied it and often tried to build machines to handle it in acoustic way. Most of the Information in digital world is accessible to a few who can read or understand a particular language. Language technologies can provide solutions in the form of natural interfaces so the digital content can reach to the masses and facilitate the exchange of information across different people speaking different languages. These technologies play a crucial role in multi-lingual societies such as India which has about 1652 dialects/native languages.

A text to speech converter convert's normal language text into speech. Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

Here question arises that whether machine or simply computer can perform same task of text to speech conversion? Answer is not that much easily as human can. The machine has to follow some procedure which is divided in basic two steps: I : character recognition for this one can OCR that is optical character recognition method . Next step is TTS that is Text to speech conversion in this we have to convert recognized text from OCR into .wav file or simply in speech file.

II. OPTICAL CHARACTER RECOGNITION

Optical character recognition usually abbreviated OCR, is the mechanical or electronic translation of images of and written, type written or printed text. optical character recognition belongs to the family of techniques performing automatic identification. For character recognition authors used different methods that are stated below.

Rama mohan babu, p. Srimaiyee, a. Srikrishna they used Characters in a text are of different shapes and structures. Text extraction may employ binarization or directly process the original image it consist a survey of existing techniques for page layout an analysis. Mathematical morphology is a topological and geometrical based approach for image analysis. It provides powerful tools for extracting geometrical structures and representing shapes in many applications. Morphological feature extraction techniques have been efficiently applied to character recognition and document analysis, especially if dedicated hardware is used.

They proposed an algorithm for text extraction based on morphological operations.
M. Nagamani, S.Manoj Kumar, S.Uday Bhaskar used OCR which is the acronym for Optical Character Recognition. This technology allows a Machine to automatically recognize a character through an optical mechanism.OCR is the process of translating scanned images of typewritten text into machine-editable information. If we read a page in language other than our own, we may recognize the various characters, but be unable to recognize words. However, on the same page; we are usually able to

interpret numerical statements-the symbol for numbers are universally used.

In this paper at first input image is converted to binary matrix. Then, the binary matrix is divided into individual digits. And each individual digit is converted into frinz distance matrix. To find frinz distance, the simplest method is frinz distance method. By using this method, find the distance between the pixels in binary image. After finding the distance, the distance is replaced by the values 0 or 225 in pixel position. Individual digit of the frinz distance matrix is subtracted with the each data base template. Each subtracted matrix elements are summed. After finding the sum, Minimum sum is then given to the speech synthesis.

Alexandre Trilla used Text Normalization which Adapts the input text so as to be synthesized. Then The sentence segmentation is achieved though dealing with punctuation marks after that The tokenization is used to separates the units that build up a piece of text.

At last non-standard words such as certain abbreviations (Mr., Dr., etc.), date constructs, phone numbers, acronyms or email and URL addresses need to be expanded into more tokens (units) in order to be synthesized correctly.

D.Sasirekha, E.Chandra proposed, the Text analysis part which is preprocessing part which analyse the input text and organize into manageable list of words. Text detection is used to localize the text areas from any kind of printed documents then Text Normalization is performed so that the transformation of text to pronounceable form. After that Linearization is used which is the process of giving a hyper text link to give the user a quick overview of the page.

Bhushan Sonawane, Kiran Patil, Nikhil Pathak, Ram Gamane used the Microsoft Office Document Imaging OCR technology to extract text tokens, prototypes and templates. Then they preformed following processes,

A.Extracted Text From Image
Webcam captured image will be processed by MODI used in “Third Eye : An Image Explorer”. An text will be extracted from image & kept in separate text file same name as image file name.
B. Text Analysis & Text Detection
Text analysis is mainly concern with analysis of extracted text from image which is in text file. Organize and maintain them into a list of words. This list contains abbreviation, numbers, and acronyms & converts them into a full line when needed. Text Detection is a process of identifying preciously where it is located in that page image.

C. Text Transformation
It is normalization of text to pronounceable from. It pronounces line by line words take pause when space is detected between words It reads the text according to the punctuation rules, accent marts & stop words much similar as human being.

Priyanka Jose , Govindaru V states at First discuss about the OCR system partially developed by C-DIT Trivandrum. The process of OCR involves several steps includes scanning pre-processing, feature extraction and post-processing.

A. Scanning :
Text digitization is a process to convert the image into proper digital image. This can be performed either by a flat-bed scanner or a hand-held scanner. Scanned image has a resolution level typically 300-1000 dot per inch for better accuracy of text extraction and saves it in preferably TIF,JPG and GIF format.

B. Pre-processing :
Pre-processing consists of a number of preliminary steps to make the raw data usable for recognizer[. Firstly the scanned image is converted to gray scale image by binarization method. sometimes skew detection and correction method is necessary to digitized image to make text lines horizontal. The noise free image is passed to the segmentation step, where the image is segmented in to characters. Various segmentation processes are explained in. It is the most important aspect of pre-processing stage.

C. Feature extraction and Classification:
All characters will be divided into geometric elements like lines, arc and circles and compare the combination of these elements with stored combination of known characters. Common feature extraction and classification method is explained in .

D. post-processing:
Remaining step is post-processing in reorganization. It include spell checking, error checking and text editing etc, when the recognized character does not match with the original one or cannot be recognized from the original one. Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J. Wu, Andrew Y. Ng proposed system proceeds in several stages:

1) unsupervised feature learning algorithm to a set of image patches harvested from the training data to learn a bank of image features.
2) Evaluation of the features convolutionally over the training images. Reduce the number of features using spatial pooling.
3) Train a linear classifier for either text detection or character recognition these stages are briefly explained here,

A. Feature learning:
The key component of our system is the application of an unsupervised learning algorithm to generate the features used for classification. Many choices of unsupervised learning algorithm are available for this purpose, such as autoencoders, RBMs, and sparse coding .Here, however, they use a variant of K-means clustering that has been shown to yield results comparable to other methods while also being much simpler and faster.

B. Feature extraction:
For this both detector and character classifier consider 32-by-32 pixel images. To compute the feature representation of the 32-by-32 image, they compute the representation described above for every 8-by-8 sub-patch of the input, yielding a 25-by-25-by-d representation. Formally,they will let z(ij) 2 Rd be the representation of the 8-by-8 patch located at positioni, j within the input image. At this stage, it is necessary to reduce the dimensionality of the representation before classification. A common way to do this is with spatial pooling where they combine the responses of a feature at multiple locations into a single feature. In our system, we use average pooling: we sum up the vectors z(ij) over 9 blocks in a 3-by-3 grid over the image, yielding a final feature vector with 9d features for this image.

C. Text detector training: For text detection, they train a binary classifier that aims to distinguish 32-by-32 windows that contain text from windows that do not. They build a training set for this classifier by extracting 32-by-32 windows from the ICDAR 2003 training dataset, using the word bounding boxes to decide whether a window is text or non-text.5 With this procedure, they harvest a set of 60000 32-by-32 windows for training (30000 positive, 30000 negative).after that use the feature extraction method described above to convert each image into a 9d-dimensional feature vector. These feature vectors and the ground-truth “text” and “not text” labels acquired from the bounding boxes are then used to train a linear SVM. They later used feature extractor and the trained classifier for detection in the usual “sliding window” fashion.

D. Character classifier training: For character classification, they used a fixed-sized input image of 32-by-32 pixels, which is applied to the character images in a set of labeled train and test datasets.6 However, since they can produce large numbers of features using the feature learning approach above, over-fitting becomes a serious problem when training from the (relatively) small character datasets currently in use. To help mitigate this problem, we have combined data from multiple sources. In particular, we have compiled our training data from the ICDAR 2003 training images, Weinman et al.’s sign reading dataset, and the English subset of the Chars74k dataset. their combined training set contains approximately 12400 labeled character image In this paper they have produced a text detection and recognition system based on a scalable feature learning algorithm and applied it to images of text in natural scenes.

A. Chauhan, Vineet Chauhan,Surendra P. Singh,Ajay K. Tomar, Himanshu Chauhan uses following procedure for text recognition purpose.

A. Text Processing:
The Text processing module consists of preprocessing and syllabication modules. The text in transliterated form is preprocessed to remove invalid characters in the text. And also, preprocessing module adds phrase break indicators to the text based on full stops and case markers. The preprocessed text is further passed on to the syllabication module.
B. Syllabication:
In this approach, the syllabication algorithm breaks a word such that there are minimum numbers of breaks in the word, as minimum number of joins will have fewer arte facts. The algorithm dynamically looks for polysyllable units making up the word, cross checks the database for availability of units, and then breaks the word accordingly.[8]

Adil Farooq, Ahmad Khalil Khan, Gulistan Raja states that, 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.[9]

Orhan Karaali, Gerald Corrigan, Noel Massey, Corey Miller, Otto Schnurr and Andrew Mackie uses the Motorola text-to-speech system which consists of three principal modules: a linguistic module, an acoustic module and a visual module. The linguistic module is responsible for generating a linguistic representation from text. The acoustic module is responsible for generating speech from the linguistic representation.

The linguistic module employs a letter-to-sound neural network and a postlexical neural network. The acoustic module employs a duration neural network and a phonetic neural network.[10] Julinda Gllavata, Ralph Ewerth and Bernd Freisleben uses, Step 1: Image Preprocessing. If the image data is not represented in YUV color space, it is converted to

this color space by means of an appropriate transformation. After that, luminance value thresholding is applied to spread luminance values throughout the image and increase the contrast between the possibly interesting regions and the rest of the image.
Step 2: Edge Detection. This step focuses the attention to areas where text may occur. They employ a simple method for converting the gray-level image into an edge image. Their algorithm is based on the fact that the character contours have high contrast to their local neighbors. As a result, all character pixels as well as some non-character pixels which also show high local color contrast are registered in the edge image. In this image, the value of each pixel of the original image is replaced by the largest difference between itself and its neighbors (in horizontal, vertical and diagonal direction). Despite its simplicity, this procedure is highly effective. Finally, the contrast between edges will be increased by means of a convolution with an appropriate mask.
Step 3: Detection of Text Regions. The horizontal projection of the edge image is analyzed in order to locate potential text areas. Since text regions show high contrast values, it is expected that they produce high peaks in horizontal projection. First, the histogram F is computed, In subsequent processing, the local maxima are calculated by the histogram determined above. Two thresholds are employed. A line of the imageis accepted as a text line candidate if either it contains a sufficient number (MinEdges) of sharp edges or the difference between the edge pixels in one line to its previous line is bigger than a threshold (MinLineDiff).
Step 4. Pre-processing
Pre-processing steps are necessary to improve the performance and make the process efficient to the time. This includes gray-scaling and binarization of image and filtering to remove noise.[11] Prof. Amit Choksi, Nihar Desai, Ajay Chauhan, Vishal Revdiwala, Prof. Kaushal Patel proposed

A. Algorithm that uses Gaussian pyramid
1. Create a Gaussian pyramid by convolving the input image with a Gaussian kernel and successively down-sample each direction by half. (Levels: 4)
2. Create directional kernels to detect edges at 0, 45, 90 and 135 orientations.
3. Convolve each image in the Gaussian pyramid with each orientation filter.
4. Combine the results of step 3 to create the Feature Map.
5. Dilate the resultant image using a sufficiently large structuring element (7x7) to cluster candidate text regions together.
6. Create final output image with text in white pixels against a plain black background.

B. Algorithm that uses Prewitt edge-detector
1. Convert the image into monochrome image by thresholding.
2. Filter the image for removing noise. Use Gaussian low-pass filter.
3. Apply Prewitt edge-detector to the filtered image.
4. Apply proper morphological operations, i.e. dilation to make clusters of text regions.
5. Multiply the resultant image with input black and white image to get text in contrast with Plain background.

c. Edge-Detection
Edges are those places in an image that correspond to object boundaries. Edges are pixels where image brightness changes abruptly. Specifically in text data probably more edges are present than non-text areas. For example, Letters „E□, „Z□, „H□,□A□ etc. are having horizontal and/or vertical edges. If to detect these edges, there may be likelihood of other letters or words around (because words are usually grouped) Thus, the text region is detected. We chose Prewitt amongst several edge-detectors available like Sobel, canny and Roberts. Choice of Prewitt is quite empirical. Prewitt edge-detector detects horizontal and vertical edges in an image and combines them to give resultant image.

C. Morphological Operations
After detecting text region(s), a cluster of it is created such that the all letters are covered. Morphological dilation is used for this purpose as dilation adds pixels to the boundaries of objects in an image thereby thickening that object. Measure of thickness is defined by the type and size structuring element. Proper sized structuring element should be chosen such that least non-text area should be clustered within. Here, structuring element „disk□ with size 9 (a disk of radius 9) is used. To remove non-text objects significantly, morphological opening operation is used. Opening operation is erosion followed by dilation. It is performed to remove objects of specific size from image. This size is again determined by structuring element. After performing such operations, the resultant image holds clusters of text regions having pixel value 1 (white).

D. Character extraction
This step refers to identify the characters as they are in original image. This is done by multiplying resultant image with binary converted original image. In this operation, pixels having value 1 (i.e. text) are recovered as same in original image and pixels having value 0 are present as background. However, the final image may contain some non-text part, extent of which is measured by precision rate. Final result is the white text in black background or vice versa, dependent on the original image.

In order to add with that Prewitt edge-detector algorithm is more robust since it gives higher recall rate as compared Gaussian edge-detector algorithm. The edge based algorithm is also able to give better

results in case of lighting variance compare to Gaussian edge detection method and also on illuminated images[12]

III. SPEECH SYNTHESIS

A Text-To-Speech (TTS) synthesizer is a computer-based system that should be able to read any text aloud, when it is directly introduced in the computer by an operator.

IV. WHAT IS SPEECH SYNTHESIS?

A Text-To-Speech (TTS) synthesizer is a computer-based system that should be able to read any text aloud, when it is directly introduced in the computer by an operator. It is more suitable to define Text-To-Speech or speech synthesis as an automatic production of speech, by 'grapheme to phoneme' transcription.

A grapheme is the smallest distinguishing unit in a written language. It does not carry meaning by itself. Graphemes include alphabetic letters, numerical digits, punctuation marks, and the individual symbols of any of the world's writing systems. A phoneme is "the smallest segmental unit of sound employed to form meaningful utterances"

The first task faced by any TTS system is the conversion of input text into linguistic representation, usually called text-to-phonetic or grapheme to phoneme conversion. This consist different methods followed by different author that are summarized as follows:

A. Chauhan,Vineet Chauhan,Surendra P. Singh, Ajay K. Tomar, Himanshu Chauhan uses

The basic types of synthesis system the following are:
• Formant
• Concatenated
• Prerecorded

A. Concatenative Synthesis:
Concatenative synthesis is based on the concatenation (or stringing together) of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech.

There are three main sub-types of concatenative synthesis. 1. Unit Selection Synthesis: Unit selection synthesis uses large databases of recorded speech. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, diphones, half-phones, syllables, morphemes, words, phrases, and sentences 2. Diphone Synthesis: Diphone synthesis uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language.

B. Formant Synthesis:
Formant synthesis does not use human speech samples at runtime. Instead, the synthesized speech output is created using additive synthesis and an acoustic model (physical modelling synthesis) Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. This method is sometimes called rules-based synthesis;

C. Prerecorded Synthesis:
In prerecorded synthesis record large paragraphs of Hindi words (commonly used Hindi vocabulary) in a continuous rhythm with small gap between two successive words in form of a silence and save them as sound files on the database.

Orhan Karaali, Gerald Corrigan, Noel Massey, Corey Miller, Otto Schnurr and Andrew Mackie proposedsystem for speech synthesis which consist Vocoder: The phonetic neural network is not trained to generate speech directly. Instead, it is trained to produce a sequence of acoustic descriptions of ten-millisecond frames of speech. These are then synthesized using a vocoder.

Tapas Kumar Patra Biplab Patra,Puspanjali Mohapatra states, the approach used here is a concatenative one. Most high quality speech synthesizers today are concatenative synthesizers. In a concatenative system, a person records speech containing a large set of basic sound units, usually corresponding to a relatively short sequence of phonemes. The units are excised from the speech, and in most systems, the units are processed with some type of speech coding method, and the resulting templates are stored in an inventory. For synthesis of a new utterance, given a phonetic transcription, the system uses rules to select the appropriate units, extracts them from the inventory and concatenates them.

Now, the phonemes can be concatenated to give various words. Since, all these sounds (phonemes) are just column vectors, their constituent elements could be placed one after another and stored in another variable (vector). This is concatenation. This way all the words could be played by merely selecting the phonemes and placing the phoneme vectors one after another.

Pijus KASPARAITIS uses the MBROLA synthesizer based on the concatenation of diphones (original technology worked out at TCTS laboratory) was created within the framework of the MBROLA project. It takes a list of phonemes as input, together with durations (in milliseconds) of phonemes and a piecewise linear description of pitch. The latter are defined as a set of pairs of numbers where the first denotes the distance from the beginning of the sound

in per cent and the second the height of the fundamental frequency (in hertz)

The synthesizer produces speech signals of 16 bits, the sampling rate of which is determined by the sampling rate of the diphone database used. A text to speech (TTS) synthesizer is a system that can read text aloud automatically, which is extracted from Optical Character Recognition (OCR). A speech synthesizer can be implemented by both hardware and software. Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer. A text-to-speech (TTS) system converts normal language text into speech. A synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output Silvio Ferreira, C´eline Thillou}, Bernard Gosselin states that, a Text-To-Speech (TTS) synthesizer is a computer-based system that should be able to read any text aloud. In this definition, TTS means automatic production of speech, through a grapheme-to-phoneme transcription of the sentences to utter. For performing the grapheme-to-phoneme transcription, the TTS synthesizer involves a Natural Language Processing module that analyzes the text. The transcription is then processed by a Digital Signal Processing module, which generates the corresponding speech signal. Bhushan Sonawane1, Kiran Patil2, Nikhil Pathak3, Ram Gamane uses, Acoustic Processing :Acoustic means “relating to sound ” or hearings same as aural. It presents voice characteristics of person. There are 3 types of acoustic synthesized available:

1) Concatenative Synthesis.
2) Formant Synthesis.
3) Articulatory Synthesis.

1. Concatenative Synthesis:
Concatenative Synthesis is pre-recorded human voice, in this process a database is needed having all the pre-recorded words. Main advantage is natural voice. Drawback is the using & developing of large database.

2. Format Synthesis:
Format Synthesized speech does not have any database of speech samples. So the speech is artificial & robotic.

3. Articulatory Synthesis:
Speech organs are called are called Articulators. It produces a complete synthetic output, based on mathematical models.

D.Sasirekha, E.Chandra proposed, Acoustic Processing in which the speech will be spoken according to the voice characteristics of a person, There are three type of Acoustic synthesing available (i).Concatenative Synthesis (ii).Formant Synthesis (iii).Articulatory Synthesis The concatenation of prerecorded human voice is called Concatenative synthesis, in this process a database is needed having all the prerecorded words.

The natural sounding speech is the main advantage and the main drawback is the using and developing of large database. Formant-synthesized speech can be constantly intelligible .It does not have any database of speech samples. So the speech is artificial and robotic. 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, typically based on mathematical models.
K. Partha Sarathy, A.G.Ramakrishnan uses the concatenative text-to-speech system and discuss the issues relevant to the development of a Marathi speech synthesizer using different choice of units: Words, dip hone and trip hone as a database. Here we are using IPA method through which we could come to know that which is a previous Unit which one is a current unit and next unit.

Prof. Sheetal A. Nirve, Dr. G. S. Sable used Recognition algorithms like Neural Network. Neural network is also known as Artificial Neural Network (ANN), is an artificial intelligent system which is based on biological neural network. Neural networks able to be trained to perform a particular function by adjusting the values of the connections (weight) between these elements.

Neural network is adjusted and trained in order the particular input leads to a specific target output. The network is adjusted, based on a comparison of the output and the target until the network output is matched the target. Now a days, neural network can be trained to solve many difficult problems faced by human being and computer.

The TTS system used is unit selection based concatenative speech synthesizer, where a speech unit is selected from the database based on its phonetic and prosodic context.

CONCLUSION:

In this paper we discussed character recognition and speech synthesis techniques which are very useful to perform the task text to speech conversion. As stated TTS is divided into two sub problems character recognition and speech. To get best speech synthesis rate, database of system should be large. So, there is scope to increase the database of proposed system. Techniques to solve these problems have also summarized with help of different papers.

REFERENCES:

[1] G. RAM A MOHAN BABU, P. SRIMAIYEE, A. SRIKRISHNA, "TEXT EXTRACTION FROM ETROGENOUS IMAGESUSING MATHEMATICAL MORPHOLOGY",Journal of Theoretical and Applied Information Technology 2005-2010.

[2] M. Nagamani, S.Manoj Kumar, S.Uday Bhaskar, " Image to Speech Conversion System for Telugu Language", International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 6, November 2013.

[3] Alexandre Trilla, "Natural Language Processing techniques in Text-To-Speech synthesis and Automatic Speech Recognition",Departament de Tecnologies M`edia Enginyeria i Arquitectura La Salle (Universitat Ramon Llull), Barcelona, Spain2009.

[4] D.Sasirekha, E.Chandra, "TEXT TO SPEECH: A SIMPLE TUTORIAL"International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[5] Bhushan Sonawane,Kiran Patil, Nikhil Pathak, Ram Gamane, "Third Eye : An Image Explorer", International Journal of Emerging Technology [5]and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 4, April 2013)

[6] Priyanka Jose , Govindaru V, " Malayalam Text-to-Speech" International Journal of Engineering and Technical Research (IJETR) ISSN: 2321-0869, Volume-1, Issue-3, May 2013.

[7] Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J. Wu, Andrew Y. Ng, " Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning",Computer Science Department Stanford University USA

[8] A. Chauhan, Vineet Chauhan, Surendra P. Singh, Ajay K. Tomar, 5Himanshu Chauhan, "A Text to Speech System for Hindi using English Language", 322 International Journal of Computer Science and Technology IJCST Vol. 2, Issue 3, September 2011.

[9] Adil Farooq, Ahmad Khalil Khan, Gulistan Raja, "Implementation of a Speech Based Interface System for Visually Impaired Persons",Department of Electrical Engineering, UET Taxila, Pakistan. Life Science Journal 2013.

[10] Orhan Karaali, Gerald Corrigan, Noel Massey, Corey Miller, Otto Schnurr and Andrew Mackie, " A HIGH QUALITY TEXT-TO-SPEECH SYSTEM COMPOSED OF MULTIPLE NEURAL NETWORKS", IEEE International Conference on Acoustics, Speech and Signal Processing, Invited paper, Seattle, May 1998.

[11] Julinda Gllavata1, Ralph Ewerth1 and Bernd Freisleben, "A Robust Algorithm for Text Detection in Images"

[12] Prof. Amit Choksi1, Nihar Desai2, Ajay Chauhan3, Vishal Revdiwala4, Prof. Kaushal Patel5, "Text Extraction from Natural Scene Images using Prewitt Edge Detection Method", International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 12, December 2013 ISSN: 2277 128X.

[13] Tapas Kumar Patra , Biplab Patra ,Puspanjali Mohapatra "Text to Speech Conversion with Phonematic Concatenation"International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 2 Issue 5 (September 2012) ISSN:2249-7838 .

[14] Pijus KASPARAITIS , "Diphone Databases for Lithuanian Text-to-Speech Synthesis", INFORMATICA, 2005, Vol. 16, No. 2, 193–202 193, 2005 Institute of Mathematics and Informatics, Vilnius.

[15] K. Partha Sarathy, A.G.Ramakrishnan, "TEXT TO SPEECH SYNTHESIS SYSTEM FOR MOBILE APPLICATIONS",Department of Electrical Engineering, Indian Institute of Science, Bangalore, India.

[16] Prof. Sheetal A. Nirve,Dr. G. S. Sable,"Optical character recognition for printed text in Devanagari using ANFIS", International Journal of Scientific & Engineering Research, Volume 4, Issue 10, October-2013 236 ISSN 2229-5518 IJSER © 2013.

★★★