A Survey on Digitization Using Image to Text to Speech Converter

Ankur Karmacharya[1], Shreya Pradhan[2], Snehashish Kumar[3], Ruchi Rathore[4], Saritha A N[5]

[1,2,3,4] Final year B.E students, [5] Assistant Professor
Department of Computer Science & Engineering, BMSCE, Bangalore

***ABSTRACT*: An image to text to speech converter is an application that recognizes text in scene images and synthesizes the text recognized into speech. The image is preprocessed and rectified before text recognition. The recognized text is converted to speech using Natural Language Processing (NLP) and Digital Signal Processing (DSP). Image to text to speech converter which will be useful in digitization as well as for differently abled people with poor eyesight, dyslexia etc.**

KEYWORDS: *Scene image, Natural Language Processing, Digital Signal Processing*

1. INTRODUCTION

The effort to convert printed text to document format dates back to the early 1900s. It involved telegraphy. Emanuel Goldberg developed a machine which read characters and converted them to standard telegraphic code. Similarly, the effort for speech synthesis commenced in the 18th century, even before the existence of electronic devices. There have been tremendous improvements in these fields compared to the early times. However, there are still lots of challenges making an image to text to speech conversion a work in progress. The wide variety of available fonts and the emergence of natural scene image recognition are the current challenges.

The first phase in the converter is recognition of text in scene images. Scene images require correction and preprocessing before actual text recognition can be performed. Computing under handheld devices involves a number of challenges. Because of the non-contact nature of digital cameras attached to handheld devices, acquired images very often suffer from skew and perspective distortion [4]. Perspective distortion is the transformation of features of the image which differs considerably when taken from different angles.

Corrections to scene images include text region segmentation, skew correction and baseline [3] detection. After appropriate corrections have been made, preprocessing needs to be done. Preprocessing steps include RGB to grayscale conversion, contrast adjustment and adaptive threshold. After appropriate preprocessing, character recognition is performed. Text recognition process in scene images is depicted in the figure below:

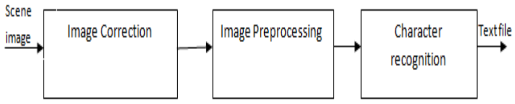

Fig 1: Basic processes involved in text recognition in images

The recognized text is written into a text file. This text file is the input for the speech synthesizer. The basic components of the speech synthesizer are Natural Language Processing (NLP) and Digital Signal Processing (DSP). NLP produces a phonetic description of the text together with its prosody. DSP receives symbolic information from the NLP and produces intelligible speech. Speech synthesis process is depicted in the figure below:

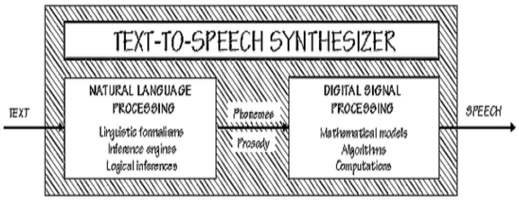

Fig 2: Basic Component of speech synthesizer

2. LITERATURE SURVEY

World without communication is desecrate. Beside different mode of communication such as art, music or even printed materials, speech has been the most primary and effective mode of communication. Many machines are built to improve the communication access but communication for visually impaired people is a challenge if it is through printed materials. Text to speech converter can be used to aid visually impaired people.

Scene images require numerous corrections. Prior to corrections, the image needs to be processed. The image is converted to grayscale. A gray scale image is an image in which the pixels are differentiated by how much light they emit. Thus each pixel is a different shade of gray. The standard formula employed in transforming an RGBimage into gray scale image is:
$0.2989 * R + 0.587 * G + 0.1140 * B$ [3].Gray scale conversion makes images simpler for further processing as it removes colors, which is irrelevant data.

The contrast of grayscale image needs to be adjusted. Contrast is difference in luminance or color that makes an object distinguishable. It is determined by differences in color and brightness of an object and other objects within the same field of view.By setting a suitable contrast, the characters in text regions can be distinguished from their background.

After the effective contrast of grayscale image has been chosen, uniformity in pixel values is achieved by Binarization. It allows only two values for each pixel in an image: black or white. In order to perform binarization, a threshold value must be chosen such that pixels having values greater than the threshold value will be classified as white while those having values lesser than the threshold value will be classified as black. For scene images local binarization techniques must be used because factors such as illumination difference. Several methods are available. In Niblack method, threshold value depends upon the local mean and standard deviation of window area. Sauvola method is a modified form of Niblack method. It gives more performance in conditions such as light variation[8]. Binarization of an image makes computations less expensive andeasier to process, as the image can have only two values, black and white.

Black and white is used as they contradict each other and are easily differentiable, which is helpful to segregate foreground from background.This is the last step of processing after grayscale conversion and contrast adjustment. Several corrections need to be performed on the preprocessed image before actual character recognition can be performed as it may contain extraneous data.

Scene images contain many irrelevant regions such as diagrams. The text regions need to be extracted. Text regions can be identified by block division of images and analysis of features such as dimensions, aspect ratio, information pixel density, region area, coverage ratio, histogram[4] etc. Other discriminant functions can be applied for scene image text region extraction after using DCT based high pass filter to remove constant background[10]. DCT is an image compression technique wherein constant irrelevant background can be removed. Kohei et. al[10] used connected component labeling method, morphology erosion filter to extract text from complex background. Shyamaet.[4] al used color based segmentation to extract text from any type of camera grabbed frame image or video. Shivakumara et. al[7] proposed a new method based on Maximum Color Difference (MCD) andBoundary Growing Method (BGM) for detection of multi oriented handwritten scene text from video.

The extracted text regions from scene images usually suffer from skew and perspective distortion[4]. Skew correction can be done by bounding text regions with virtual rectangle and utilizing profiles[4]. Some techniques apply Eigen point clustering algorithm[9]. Once skew correction has been applied character recognition is applied.Document text extraction has been extensively and successfully performed using Optical Character Recognition (OCR). The basic processes involved in Optical Character Recognition are shown in the figure below:

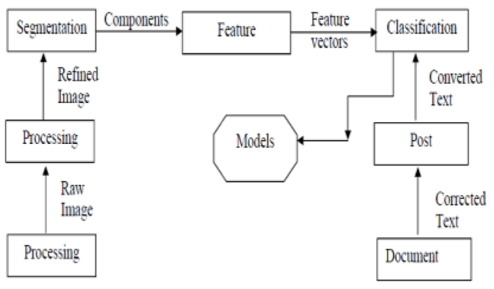

Fig 3: Basic working of an OCR system

Tesseract OCR is an open source engine developed by HP between 1984 and 1994[2]. Currently, the project is sponsored by Google.

The task of recognizing text and digits in a controlled setting has achieved high accuracy[3]. However, detection and recognition of characters in a scene image still remains an area of research. Clever techniques have been employed for scene images. Some solutions use simple classifiers, which are trained on specific hand-coded features[3] while some use a combination of different algorithms[3][2]. Some systems employ flexible technique to learn necessary information from labeled data with minimum prior knowledge. Multi layered neural network architecture has been implemented in character recognition quite successfully[4]. Coates el al [9] extracted local features of character patches from an unsupervised learning technique associates with a variant of K-means clustering,and grouped them by cascading sub patch features. S.M. Lucas et al[8] carried out a complete performance evaluationof scene text character recognition to design a distinctive feature representation of scene text character structure. Epshtein et al[9] described a real time scene text localization and recognition method based on external regions. Mishra et al[10] implemented conditional random field to combine bottom-up character recognition and top down word recognition. The recognized characters are stored in a text file. The text can now be synthesized to speech.

Text to speech synthesizers broadly consists of two major modules: Natural Language Processing (NLP) and Digital Signal Processing (DSP). NLP produces a phonetic description of the text together with its prosody. DSP receives symbolic information from the NLP and produces intelligible speech. The first step in an NLP is token to word conversion. Tokens are numbers and shorthand notations used in text. The token to word conversion creates an orthographic form of the token[3]. The next step is application of pronunciation rules. In applications like public announcements in bus and train stations, counters in banks etc. where the extent of vocabulary used is limited, the pronunciation of words to be used can be stored in a database. Subsequently, generation of speech becomes a simple matter of identifying words and generating speech, as is present in the database. For general application, the extent of vocabulary used is very large. One possible implementation could be to store all words present in a dictionary and their phonetic descriptions in the database. The problem with this method is the necessity of a huge database. Also it doesn't address the issue of prosody. An alternative approach is a rule based approach where the rules for generation of phones are described, which is then followed to generate the pronunciation of words. This way we don't require a huge database. It also accounts for prosody generation. However, this technique fails to address the anomalies such as silent letters. To counter this problem, words having anomalies in their pronunciation can be stored separately. After phonetic description has been generated, the next step is to generate prosody.

Prosody is the related to the pronunciation of text in the context of the sentence it is being used. The degree of naturalness of a TTS system is dependent on prosodic factors like intonation modeling (phrasing and accentuation), amplitude modeling and duration modeling (including the duration of sound and the duration of pauses, which determines the length of the syllable and the tempos of the speech)[3]. Upon availability of phonetic description from the NLP, a DSP then synthesizes speech. There are several methods to implement DSP for speech synthesis. One popular method is Fourier transform. Fourier transform characterizes any signal as a sum of set of sinusoids. The output signal will be similar to the input signal with shifted amplitudes and phases by a particular frequency that is dependent on the particular system. Once we measure how all relevant frequencies are modified by the system, it becomes easy to estimate system's output based on its input using the Fourier transform. The phonetic description will serve as the system's input and output will be intelligible speech. Thus, an image with text after being preprocessed through several enhancement and corrective stages is recognized and placed into a text file, which is then converted to intelligible speech through application of Natural Language Processing and Digital Signal Processing methods.

3. CONCLUSION

While image to text conversion in printed documents has been extensively researched and implemented with almost full accuracy, the field of text recognition in scene images still remains a topic of active research. The process of converting scene image to text to speech consists of a number of sub processes. Each sub process has numerous implementations in existence. Each implementation has its merits and demerits. Further, many different combinations of sub processes exist, each combination having its own share of efficiencies and inefficiencies. This converter has a wide scope of application provided that it is sufficiently accurate.

4. ACKNOWLEDGEMENT

The work reported in this paper is supported by the college through the TECHNICAL EDUCATIONAL QUALITY IMPROVEMENT PROGRAMME [TEQIP-II] of the MHRD, Government of India.

REFERENCES

[1] M. Prabaharan, K. Radha, "Text Extraction from Natural Scene Images and Conversion to Audio in Smart Phone Applications", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2015.

[2] Mrunmayee Patil, RameshKagalkar, "A Review on Conversion of Image to Text As Well As Speech Using Edge Detection and Image Segmentation", International Journal of Science and Research, Volume 3, Issue 11, November 2014.

[3] Shivakumara, K.Kalaivani, R.Praveena, V.Anjalipriya, R.Srimeena, "REAL TIME IMPLEMENTATION OF IMAGE RECOGNITION AND TEXT TO SPEECH CONVERSION", International Journal of Advanced Engineering Research and Technology (IJAERT) Volume 2 Issue 6, September 2014.

[4]Ayatullah Faruk Mollah, Nabamita Majumder, Subhadip Basu, Mita Nasipuri,
Shyamay "Design of an Optical Character Recognition System for Camera-based Handheld Devices", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2014.

[5] Sayali Shukla, Ashwini Sonawane, Vrushali Topale, Pooja Tiwari, "Improving Degraded Document Images Using Binarization Technique",International Journal Of Scientific & Technology Research Volume 3, Issue 5, May 2014.

[6] ItunuoluwaIsewon, JeliliOyelade, OlufunkeOladipupo, "Design and Implementation of Text To Speech Conversion for Visually Impaired People", International Journal of Applied Information Systems, Volume 7– No. 2, April 2014.

[7] Shivakumara, "Binarization techniques used for Grey Scale Images", International Journal of Computer Application, 2013.

[8] A. Coates et al., "Text detection and character recognition in scene images with unsupervised feature learning", in Proc. ICDAR, pp. 440–445, 2011.

[9] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform", in Proc. CVPR, pp. 2963–2970, 2010.

[10] Kohei, Angadi, S.A., "Text region extraction from low resolution natural scene images using texture features", Advance Computing Conference (IACC), IEEE 2nd International Journal, 2010.

[11] Y. Pan, X. Hou, and C. Liu, "Text localization in natural scene images based on conditional random field", in International Conference on Document Analysis and Recognition, 2009.

[12] J. Zhang and R. Kasturi, "Extraction of text objects in video documents: Recent progress", in Proc. 8th IAPR Int. Workshop DAS, pp. 5–17, 2008.

[13] Y. Pan, X. Hou, and C. Liu, "A robust system to detect and localize texts in natural scene images", in International Workshop on Document Analysis Systems, 2008.

[14] J. J. Weinman, E. Learned-Miller, and A. R. Hanson, "A discriminative semi-Markov model for robust scene text

recognition", in Proc. IAPR International Conference on Pattern Recognition, Dec. 2008.

[15] Z. Saidane and C. Garcia, "Automatic scene text recognition using a convolutional neural network", in Workshop on Camera-Based Document Analysis and Recognition, 2007.

[16] X. Chen and A. Yuille, "Detecting and reading text in natural scenes," in Computer Vision and Pattern Recognition, vol. 2, 2004.

[17] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions", in Proc. Int. Conf. Document Anal. Recognit., pp. 682–687, 2003.

[18] J. Sauvola, , M. PietikaKinen, "Adaptive document image binarization" , the journal of pattern recognition society, 21 January 1999.

[19] Yang Cao, Heng Li, "Skew Detection and Correction in Document Images Based on Straight-Line Fitting", February 1998.

[20] Text-to-speech technology: In Linguistic Language Technology Website. Retrieved February 21st,2014, from http://www.linguatec.net/products/tts/information/technology .