A Review on the Various Techniques used for Conversion of Text in an Image into Speech

Amogh Bhatnagar[1], Mahek Chhabra[2], Anirudh S K[3], Vipul Gohil[4]

[1,2,3]Student, Btech EXTC, Mukesh Patel School of Technology Management and Engineering, Mumbai India,
[4]Assistant Professor, Dept. of EXTC, Mukesh Patel School of Technology Management and Engineering, Mumbai India

---

Abstract - *The main problem in communication is the language bias between communicators. The use of a translator in foreign lands is highly helpful, and using a system for the translation of a text on a signboard to a language known to the user, and then to a voice, is beneficial. An image-to-speech system provides the output after undergoing four stages – Image pre-processing, Text extraction, Text translation & Text-to-Speech (TTS). Each of these four stages can be done through various techniques and methods and this paper provides a review of the various techniques used for the final procedure. The purpose of this survey is to categorize and briefly review the literature on image pre-processing, text translation and text-to-speech conversion, which comprises of 11 papers published in the last 3 years. For the reviewed techniques, some performance evaluation metrics achieved in various experiments are emphasized to help the researchers when making choices and develop new techniques.*

Key Words: **Optical Character Recognition (OCR), Peak Signal to Noise Ratio (PSNR), Statistical Machine Translation (SMT), Text-to-Speech (TTS).**

1. INTRODUCTION

A basic image-to-speech model has endless applications in today's world. The final output, speech, is possible using the combined functioning of four different processes - Image pre-processing, Text extraction, Text translation & Text-to-Speech (TTS). The failure or malfunctioning of any one these processes could result in incorrect output, thereby failing to meet the objective of the system. For each of the aforementioned, there are countless methods and techniques through which its respective output can be obtained. This paper provides a detailed study of the methods used in the processes involved, in order to get a speech output from a text present in an image.

2. IMAGE PRE-PROCESSING

It's the first step in the model, which requires the deblurring and denoising of the image, so that the text, present in the image, can be extracted with minimal amount of distortions in it. Image pre-processing addresses various techniques that include noise filtering, pseudo-colouring, sharpening, contrast and edge enhancement, contrast stretching, scaling (i.e., magnification, reduction and rotation), morphological operations, enhancement filtering, colour model conversions and histogram modifications. There are various methods used by the authors and papers [1] – [6] elucidate the various methods which can be used for the image pre-processing.

Sharma, Sharma [1] annotates the various deblurring and contrast enhancement methods on MATLAB and compares them based on the Optical Character Recognition (OCR) accuracy. The paper provides a comprehensive study of the image degradation model, deblurring techniques and contrast enhancement techniques. The authors first degrade an image using a fixed model and then apply deblurring techniques – blind & non-blind. The methods for deblurring used were – Regularized filter, Lucy Richardson, Weiner filter & Blind Deconvolution method. The deblurred image will be then enhanced in terms of contrast. The results obtained are shown in table 1 and table 2.

Hong ZHANG, Qi GE et al [2] introduces a new Point Spread Function (PSF) estimation approach for tackling the distortions and degradations caused by the atmospheric disturbances and were then applied for image restoration. The PSF is estimated via an isosceles model which is proposed to approximate one component of the original image's Fourier amplitude. On degraded image restoration, the short-exposure image frames are transformed into a single Long-exposure image, and then by the use of PSF and Weiner Filter the image is restored. There are two images that have been taken for the purpose – 'Aerial view' & 'Bridge', shown in Fig. 1 below. Gray Mean Grads (GMG) and Laplacian Sum (LS) are adopted to evaluate the results objectively. GMG reflects the contrast level of the image. Similarly, LS is able to reflect both the profile and the detail of the images. These results are shown in Table 3 below.

Zhang and Hirakawa [3] have put forward a method that extends Double Discrete Wavelet Transform (DDWT) deblurring to recover an image with minimal sensor noise and sharp blur boundaries due to object movement and an estimate of spatially varying blur kernel from a noisy and motion blurred image. The method put forward in this paper

overcomes the problem of the scattered representation of DDWT—the generative version of the blur which helps in the Wavelet Transform analysis of the input blurred image—and brings about the viewpoint (Bayesian) of the process which should model the distribution of latent sharp wavelet coefficient and the likelihood function (which makes the noise handling explicit) prior to the process. Blur kernel

Table -1: Analysis of deblurring techniques

Name of Deblurring method	Type of method	Required Information	Results
Regularized filter method	Non-blind method	Knowledge about PSF and information about noise	It provides the best results in comparison to Lucy Richardson and Weiner method.
Lucy-Richardson method	Non-blind method	Knowledge about PSF and information about noise	It provides satisfactory results but ringing effect is Prominent.
Weiner filter method	Non-blind method	Knowledge about PSF and information about noise	It provides good results only when the knowledge about noise is known.
Blind Convolution method	Blind method	No knowledge about PSF and information about noise	This helps us in guessing the PSF and returns both the de-blurred image as well as the estimated PSF with which the image was blurred.

Table -2: Analysis of contrast enhancement techniques

Name of the technique	Results	Outcome
Histogram Equalization	It does not provide good results for the tested image as the image gets darkened.	Not fit for use in the case of text images.
Imadjust	It provides good results for the tested image and improves the contrast of the image.	Fit for use in case of text Images.

Fig -1. Bridge and Aerial view images

Table -2: GMG of the degraded and restored images

GMG	Degraded image	Algorithm used earlier	Algorithm in this paper
Ariel View	6.0297	12.0813	12.2454
Bridge	2.8476	7.3231	7.6243

LS of the degraded and restored images

LS	Degraded image	Algorithm used earlier	Algorithm in this paper
Ariel view	19.1108	37.4703	50.0790
Bridge	11.0164	20.8380	34.6184

Estimation with numerical integration is used as an algorithm for the Maximum Likelihood Estimation (MLE). The results are shown in Fig. 2.

Choudhary, Sa et al [4] proposes an image restoration scheme, for images distorted due to unpredicted movement or out-of-focus blur combined with Gaussian noise. The paper discusses the use of image deconvolution and image denoising which is followed by deblurring. Further noise removal is done by selecting a test pixel and iterating its neighbouring region to variants of Weiner filter. Subsequently, the neighbouring pixels of each text pixel is selected based on the properties (noise variance and uncorrelated) of additive noise. The deblurring function uses the lower bound of the regularization and its main function is to be an edge recovery constraint. The suggested framework that this paper implies has a higher Peak Signal-to-Noise Ratio (PSNR) which is better than the other methods used such as Gaussian noise filtering based on PCNN time matrix (GPTM), Spatially adaptive denoising algorithm (SADA), Directional modified sigma filter (DMSF), Image denoising by sparse 3-D transform-domain

collaborative collaborative filtering (BM3D) and Patch-based near-optimal image denoising (PLOW). The results are shown in Table 4 below.

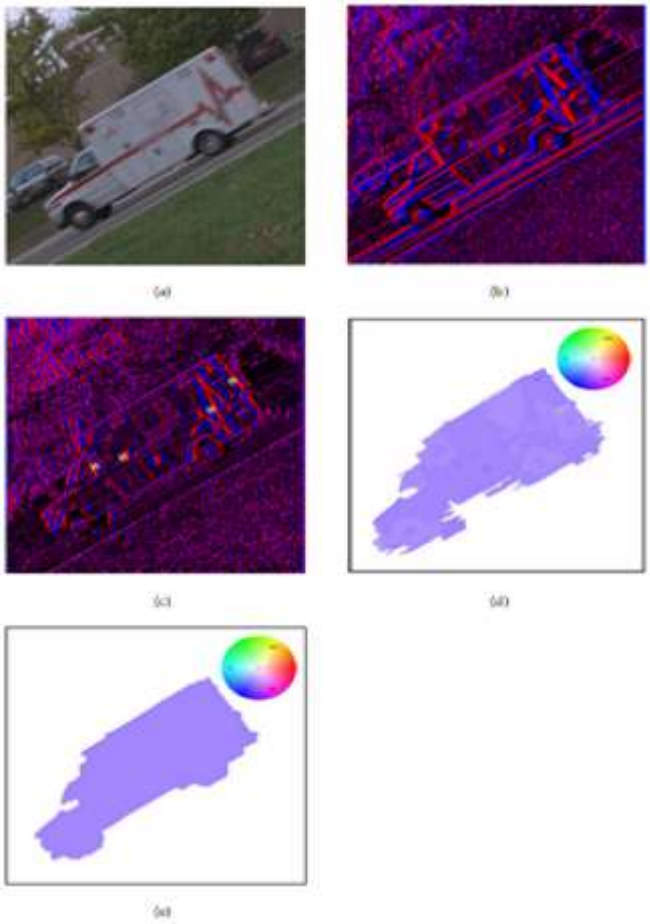

Fig -2. (a) Blurred input image y (b) DWT coefficients ω_j (c) DDWT coefficients v^{ij} (d) Intermediate blur kernel estimate (e) Final blur kernel estimate

Nithyananda, et al [5] studies the various methods used in the Histogram Equalization. This paper reviews different types of equalization methods that can be used for image enhancement and image restoration. The image is represented as a histogram diagram and then used to define the upper and lower boundaries. The classical approach is either local or global. Local operation divides the image into sub-images or masks whereas the global operation takes the image as a whole. The advantages and disadvantages of various histogram methods discussed in the paper is shown in table 5.

Periera et al [6] is a review paper in itself which analyses various techniques and methods used to enhance, segment and analyse the natural images, namely the ones used for agriculture. The aforementioned paper refers literature from different papers to detect and enhance the presence of leaves and fruits present in the image. It uses a number of techniques such as Colour Histogram, Radial Symmetry Transform, Elliptical Contour Model, Colour Mapping & Morphological Dilation etc. to segment various fruit images given to the system. For different fruits such as apple, grape, banana, mango and papaya, the authors have given different methods for feature extraction, along with their estimated percentage errors.

Table -4: PSNR (db) Comparison for restored images degraded by Gaussian noise

Image	σ_n	GPTM	SADA	DMSF	BM3D	PLOW	Proposed
Lenna	3	29.31	35.38	31.63	41.38	41.21	41.04
	6	29.03	34.13	31.19	37.49	37.4	38.11
	9	28.65	32.68	30.54	35.33	35.15	35.84
	15	27.69	29.91	29.1	32.62	32.51	32.98
	21	26.72	27.58	27.8	30.99	30.71	31.08
	24	26.3	26.6	27.2	30.38	30.04	30.32
Cameraman	3	28.79	32.16	30.05	40.87	40.49	42.37
	6	28.49	31.5	29.68	36.56	36.31	37.14
	9	28.09	30.66	29.18	34.22	33.97	34.45
	15	26.98	28.73	27.88	31.38	31.05	31.22
	21	25.9	26.87	26.63	29.73	29.32	29.77
	24	25.33	25.96	25.98	29.04	28.61	29.14
Baboon	3	24.51	28.34	25.71	38.64	38.3	38.88
	6	24.43	28.03	25.58	33.68	33.35	34.49
	9	24.29	27.53	25.37	31.04	30.75	31.04
	15	23.92	26.33	24.82	28.13	27.73	28.96
	21	23.55	25.17	24.29	26.55	25.95	26.84
	24	23.34	24.49	23.99	25.92	25.33	26.09

3. TEXT DETECTION AND EXTRACTION

The second step after the Image pre-processing is the detection and extraction of text from the image. It is the most important step of this system as incorrect text output may not give us the required speech output. The key factor that we use for text detection is the presence of feature points present in the image and later cluster them together later. The OCR software such as Tesseract can also be used for the conversion of text in images to a text file. The obtained text file is then translated into the language preferred by the user.

Lee et al [7] is a paper that outlines the Features from Accelerated Segment Test (FAST) method for the detection of the text in the images. It considers the feature points present in every image and through the help of FAST corner detection method, we can retrieve the text from the street signboard images. After the corner detection method has been used to detect the points, the system is used to cluster the feature points together and is followed by the usage of image processing techniques to minimum bounding of the text box. In order to use the detection method, the image colour space is first converted into HSV colour space. The

method used by the authors is highly effective, as compared to the previous BUCT_YST algorithm.

Table -5: Analysis of different methods of Histogram equalization

Serial Number	Algorithm	Advantages	Drawbacks
1.	HE/CHE	Good enhancement technique, based *CDF* distribution	Degraded sharpness, creates artefacts and does not preserve brightness
2.	BBHE	Preserves brightness of image	Generated image might not have natural appearance
3.	*DSIHE*	Image luminance is well maintained	Natural appearance of the Image is lost
4.	*MMBEBHE*	Provides maximum brightness preservation	Generated image has side effects, performance degraded
5.	*RMSHE*	Enhances all regions of image	Number of decomposed sub-histograms increases in a power of two.
6.	MHE	Preserves maximum brightness of the image	Performance depends on the number of sub-images.

(a) (b) (c)

Fig -3. (a) Original Image, (b) Image in HSV colour space, (c) Feature point detection

Rithika et al [8] presents a complete image-to-speech system built upon the Rasberry Pi. It involves the image being provided as the input to the system and using of the Tesseract OCR for the text detection and extraction from the image. The built-in libraries for the Rasberry Pi help greatly to use the Tesseract OCR or Google application program interface (API) as per our convenience. The installed software vastly reduces the need to manually program the system and saves a lot of processing time too.

4. TEXT TRANSLATION

The third step of this system is the translation of text into another language. This is an important step, as based on the text extracted, the translation is done. It is usually done in a language suitable to the user and is highly necessary for proper translation to take place otherwise it may cause a problem in the speech output. Papers [8] to [10] refer to this part of the procedure.

Rithika et al [8] uses the Google API software in order to translate the text extracted into any of the 52 languages provided by the software. Using the API key, it enables the user access to a large database and a variety of options to choose from. The presence of an inbuilt Grapheme-to-phoneme (G2P) system also helps the user to have a proper pronunciation of the text that has been translated from one language to another.

Yan et al [9] presents the text translation using Statistical Machine Translation (SMT) and uses the Support Vector Machines (SVM) classifier model to solve the text translation. An objective function is introduced and optimized and it is then used to select a sentence with highest probability – which serves as the main basis of SMT. This method is not to be used in this system as the optimization of the function may not result in the proper choosing of the sentence from the database of the options provided.

Chand [10] presents a review paper on the various methods that can be used for the machine translation process. Various Machine Translation processes such as Statistics based, Corpus based, Rule based and Hybrid are used for comparisons amongst them. Statistics based Machine Translation (SBMT) such as Bing, Google Translate, IMTranslator provides better results as compared to the Rule Based Machine Translation, such as Angla Bharati and Anubaad. Though the hybrid of SMT based system and Rule based system are yet to be tested, there's no real say about its efficiency. Bing clearly outperforms Google Translate and is the best possible option for Machine Translation.

5. TEXT-TO-SPEECH CONVERSION

This is the final step of the system, wherein the translated text is converted into speech. The importance of this step is to provide proper speech output which helps the user understand the translated text. The phonemic representation of the text can either be done using the inbuilt functions provided on the platform or by using any of the engines existing in the market.

Rithika et al [8] uses software named speak for the text-to-speech synthesization. Its software can be easily installed in Rasberry Pi and is used to create the phonemic representation of the text. This engine converts the text file into a free lossless audio codec file, also known as .flac file. The output of this engine is a compressed audio file, which is the final stage of the system.

Desai et al [11] provides a review of the multiple Text-to-Speech (TTS) methods existing currently – Formant based, Concatenative based and Articulatory based. Methods such as parallel, Klatt, PARCAS model, Diphones based, Vocal Tract models and Acoustic models are used to compare the methods. Generally, formant and Concatenative synthesis methods are used in present systems. The paper provides a comprehensive survey on all the techniques used in each method and provides a good comparison between each. Concatenative method is gaining popularity due to the fact that the use of discontinuity methods in concatenation points is getting effective. Even though the concatenative method provides a natural and individualistic sounding speech, it offers alterations in the quality of certain consonants, pitch and the duration of the words, especially with longer units. The collection and labelling of speech samples is often difficult and time consuming in this method. In the formant method, the quality of synthetic speech is constant but, the speech sounds are unnatural. Still, formant method is quite flexible and allows a good control over the fundamental frequency. The third method, articulatory method, is the most feasible in the theory because it has direct model of the human articulatory system itself.

6. CONCLUSIONS

An Image-to-Speech system is an expeditiously growing aspect in the present world and is increasingly playing a more important role in the way we interact with the environment around us. This image-to-speech system consists of four greatly varying stages, without whom each can't function properly. These stages, from image pre-processing to speech synthesis, each can be done using vast amounts of different procedures, which have been verbosely explained in this paper. For each stage, we've referred various technical and review papers outlining numerous methods and providing a conclusive result for the best procedure possible. Many of these methods and techniques have proven useful and efficient over the years, but however, our survey reveals that work needs to be concentrated on improving their success and efficiency. Hence, this paper can be used to select the most efficient and reliable set of methods for the creation of this prototype.

The future of the Image-to-Speech is vast, considering the fact that this is the age of globalization and involves the intermixing of people with different cultures and languages. With the help of such a system, it is invaluable to the greater understanding of various languages and can come in handy when travelling abroad. The creation of specific software, by the combination of OCR system, Machine Translation systems and API software may make this prototype an easy-to-create and user-friendly system, helping in enriching the lives of many in future.

REFERENCES

1) P. Sharma, S. Sharma: "Image Processing based Degraded Camera Captured Document Enhancement for Improved OCR Accuracy", 6th IEEE Conference – Cloud System and Big Data Engineering, 2016.

2) Hong ZHANG, Qi GE, Lu LI, Yuecheng LI, Kaiyu XI: "A New Point Spread Function Estimation Approach for Recovery of Atmospheric Turbulence Degraded Photographs", 4th International Congress on Image and Signal Processing, 2011

3) Yi Zhang, Keigo Hirakawa: "Blind Deblurring and Denoising of Images Corrupted by Unidirectional Object Motion Blur and Sensor Noise", Journal of Latex Class Files, vol. 11, no. 4, December 2012

4) SK Choudhary, PK Sa, RP Padhy, B Majhi: "A denoising inspired deblurring framework for regularized image restoration", 2016 IEEE Annual India Conference (INDICON), Bangalore, India, December 2016.

5) Nithyananda C R, Ramachandra A C, Preethi: "Review on Histogram Equalization based Image Enhancement Techniques", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016

6) Carlos S. Pereira, Raul Morais, Manuel J. C. S. Reis: "Recent Advances in Image Processing Techniques for Automated Harvesting Purposes: A Review", Intelligent Systems Conference, London, UK, September 2017,

7) Chien-Cheng Lee, Shang-Fei Shen: "Feature Point based Text Detection in Signboard Images", IEEE International Conference on Applied System Innovation (ICASI), Okinawa, Japan, 2016.

8) Rithika H., B. Nithya Santoshi: "Image Text to Speech Conversion in Desired Language by Translating with Rasberry Pi", 2016 IEEE International

Conference on Computational Intelligence and Computing Research, Chennai, India,15-17 Dec, 2016.

9) Jia Yan, Wang Chao: "English Language Statistical Machine Translation Oriented Classification Algorithm" IEEE International Conference on Intelligent Transportation, Big Data and Smart City, 2015.

10) Chand Sunita: "Empirical Survey of Machine Translation tools", 2nd IEEE International Conference on Research in Computational Intelligence and Communication Networks, 2016

11) Desai Siddhi, Varghese Jashin M., Desai Bhavik: "Survey on Various Methods of Text to Speech Synthesis", International Journal of Computer Applications, vol. 165 – No.6), May 2017