Multimode Summarized Text to Speech Conversion Application

Archit Sehgal [1], **Gitika Khanna** [2]

[1,2]Department of Computer Science, HMR Institute of Technology & Management
Hamidpur, Delhi-110036, India
Archit_150@yahoo.co.in, gitikakhanna392@gmail.com

Abstract **-** This paper draws focus towards summarizing the tremendous amount of data collected from various sources and presenting the output as speech. In recent years, huge data sets are being generated every moment and it becomes difficult to manage it. In order to extract relevant information, an innovative, efficient and real- time cost beneficial technique is required that enables users to hear the summarized content instead of reading it. This kind of application is beneficial for visually impaired and people with disabilities. Text Rank algorithm, a ranking based approach is proposed with a variation in similarity function to make summary based on the scores computed for each sentence. The summarized text is then spoken out using text-to-speech synthesizer (TTS).

Keywords **-** TextRank, PageRank, Lexemes, Image Segmentation, Character Recognition, Text-to-Speech (TTS).

I. INTRODUCTION

In our proposed work of collecting data from different sources and converting it into summarized text, we develop a cost efficient and user friendly interface. The input to the application can be an image, audio or video. While converting the input into editable text, there are various techniques used such as image processing, image segmentation [1] and edge detection. The approach direct towards format conversion, where audio, video or image data is converted into symbolic representations that fully describe the content. In case of an image, the segmented characters are obtained from preprocessing of images. It is then provided as input to the Optical Character Recognition (OCR) to obtain the converted text. In order to manage the enormous amount of information, the derived text is summarized using a graph based technique i.e. TextRank Algorithm.

The TextRank Algorithm has application in construction of meaningful summary by selecting useful paraphrases from the text available. The summarized text is then transformed into speech using Text-to-Speech Synthesizer (TTS). The whole approach is categorized into three phases which are text extraction from input, formation of summary and conversion of the same into speech.

II. LITERATURE REVIEW

Mrunmayee Patil[1] This paper tells us about an OCR system to recognize the characters from image. Edge detection and Image segmentation plays a significant role in extraction of text from image. The algorithm which can be used to summarize the extracted text works similar to PageRank Algorithm discussed in the paper for web search engines [10]. Modifications can be made to make the TextRank algorithm more effective. Sunchit Sehgal[5] This paper represents a way to make the algorithm more efficient by taking the score of the title in account. Marcia A. Bush [19] shows us the efforts put in the research of recognition of documents and their prediction models This has enabled us to analyze the signal based processes taking vocabulary, font and the sentence formation sequence into account.

III. OVERVIEW OF IMAGE ANALYSIS

Over the decades, many researchers have been looking for possible ways of retrieving data from images and video content. In a research paper, a framework was proposed that will decompose the scanned image into its constituent visual patterns and the parsed results will be converted into semantically meaningful text report. A model was also introduced where the users will send the image of their respective meter's display screen along with the kilo-watt information [2]. The information will then be processed to convert it into text.

Image analysis [3] is the extraction of information from images using different image processing approaches such as image filtering, image compression, image editing and manipulation, image preprocessing, image segmentation, feature extraction, object recognition. An Image can be considered as a matrix of square pixels arranged in the pattern of rows and columns. It can be considered as a linear sequence of characters.

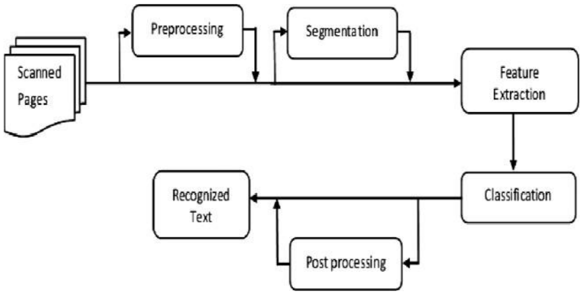

Fig 1. Conversion of Image to text using Optical Character Recognition

Fig 1. Depicts various blocks responsible for detecting text from the image. Once the image is scanned, it is preprocessed to remove any noise and is further divided into segments. Every segment has its own unique feature which must be further extracted and classified to specific groups.

Edge detection and image segmentation are important aspects in image analysis. Edge detection differentiates different regions of an image by identifying the change in gray scale and texture. Image Segmentation is another technique which divide and decompose image for further processing. It categorizes the pixels with similar gray scale values and organizes it into higher level units so that the objects become more meaningful. The proposed system will work in various phases. The input image will undergo pre-processing such as removing noise induced due to the technique applied for thresholding and improving the quality. The image will undergo image segmentation to separate non-text part present in the next step. Further, feature extraction is performed to extract preliminary features and comparing the same which are stored in the database. Sometimes, there are often error in which characters might be blurred or broken. They are processed in post-processing stage.

IV. AUDIO ANALYSIS

Conversion of real time speech to text requires special techniques as it must be quick and precise to be recognizable. In an automatic speech recognition system, the size of vocabulary affects the performance of the system. Amidst the initial process, the system learns about pattern, different speech sounds which embody the vocabulary of the application. If there is any unknown pattern, it is identified using the cluster of references. The whole approach can be categorized as phases such as analysis, feature extraction, modeling and testing. The analysis phase is used to extract information about speaker identity using vocal tract, behavior feature and excitation source. Since every speech has different characteristics which can be fetched in the feature extraction phase in order to deal with the speech signals.

V. TEXTRANK ALGORITHM

With the tremendous growth in chunks of text data, there is a need to effectively summarize it to be useful. Automatic text summarization [4] is very demanding and non-trivial task. There have been methods proposed which uses word and phrase frequency to extract salient sentences from the text. Overall, there are two different approaches for text summarization: extraction and abstraction. Extraction works by selecting the sentences from original text whereas abstraction aim at modifying the original text using advanced natural language techniques in order to generate a new brief summary. However, extractive summarization yields better results as compared to abstractive summarization because abstraction face issues such as semantic representation and natural language generation. Here, we focus on graph based TextRank algorithm to perform extractive summarization. TextRank [5] is an autonomous machine learning algorithm and is an extension of the PageRank algorithm.

VI. PROPOSED APPROACH

In our proposed approach to build this application, input can be taken in different modes such as editable, text, image, audio and video also. After the input is taken, TextRank Algorithm can be used to convert it into summarized text. The summarized text will be taken as output and convert into speech. Input processing from different modes has been discussed above. The major concern is summarizing the content efficiently and accurately. In further section we will discuss about improving the graph based technique i.e. TextRank Algorithm for accurate summarization.

For any summarizer, intermediate representation is done to express the main aspect of the text. It uses two

$$\mathbf{PR(Vi) = (1 - d) + d * X\ Vj \in In(Vi)\ PR(Vj) \mid Out(Vj)} \tag{1}$$

In order to build a connected graph, an edge is to be added between the two vertices which represents the similarity between them. The similarity depends on the words common between the two sentences which can be calculated using the similarity function. Let Si and Sj be two sentences where a sentence is represented by Ni words that forms it.

$$\mathrm{Similarity}(S, S_{i,j}) = |\boldsymbol{Wk}\ |\boldsymbol{Wk}{\in}\boldsymbol{Si}\ \&\ \boldsymbol{Wk}\ {\in}\boldsymbol{Sj}|\ \mathbf{log}(|\boldsymbol{Si}|)+\mathbf{log}(|\boldsymbol{Sj}|) \tag{2}$$

Score is accredited to each sentence depending upon the type of representation approach. In topic representation, the score depends on how well the sentence describes the topic whereas in case of indicator representation, a variety of machine learning techniques can be used in aggregating the results. In the final step, a methodology should be used which selects the best combination of sentences that maximizes the importance and minimize redundancy.

In our proposed method, TextRank algorithm is used to find the similarity between the sentences. This method describes the document as a connected graph where sentences represents the vertices and an edge indicates how similar the two sentences are. It is based on frequency of occurrence of words so any specific language processing is not required.

Consider an undirected graph, say G = (V, E), where V = set of vertices.

E = set of edges.However, the title can also play an important role in

For a given vertex Vi, let In(Vi) represent the set of vertices pointing towards the former vertices and Out(Vi) represents the set of vertices pointing to the next-inline vertices. The score can be calculated for each vertex using the formula: adding distinguishing information to elaborate meaning of the text. The similarity function can be improvised by computing the correlation between each individual sentence and title of article as well.

So, the modified similarity function between two sentences and for each individual sentence is given by:

$$\boldsymbol{Similarity}(\boldsymbol{s_i}, \boldsymbol{s_j}) = \frac{|\boldsymbol{Wk}|\boldsymbol{Wk} \in \boldsymbol{Si}\&Wk \in Sj|}{(|\boldsymbol{Si}| + |\boldsymbol{Sj}|)/_{\mathbf{2}}}$$

$$\boldsymbol{Similarity_{title}}(\boldsymbol{s_i}, \boldsymbol{s_{title}}) = \frac{|\boldsymbol{Wk}|\boldsymbol{Wk} \in \boldsymbol{Si}\&Wk \in S\boldsymbol{title}|}{(|\boldsymbol{Si}| + |\boldsymbol{Stitle}|)/_{\mathbf{2}}}$$

Therefore, the cumulative score for any sentence say S1 is given by:

$$\boldsymbol{Similarity_{title}}(\boldsymbol{s_i}, \boldsymbol{s_{title}}) + \left\{\sum_{\boldsymbol{j=i,j=2}}^{\boldsymbol{j=n}} \boldsymbol{Similarity\,Santences}\left(\boldsymbol{s_i}, \boldsymbol{s_j}\right)\right\} - \left(\boldsymbol{Similarity_{Sentances}}(\boldsymbol{s_1}, \boldsymbol{s_1})\right)$$

VII. IMPLEMENTATION AND EVALUATION

We have made this application using Apache Cordova and this application is compatible with both android and IOS. We divided the whole approach into three modules:

- Module 1: Uploading of image and processing of input text using OCR approach. The text can directly be typed in the text field. It can also be taken in the form of audio for which we used a button to record voice and then processing it using audio analysis.
- Module 2: In this module, text summarization takes place once an event is fired. Sentences are ranked and the best sentences are picked to make up a summary and be shown in the output window.
- Module 3: The text in the output window is converted into speech when a button is clicked.

We have evaluated our application by taking 3sample articles and evaluating the summary using ROUGE evaluation. ROUGE [6] is the most widely used method to evaluate the summary automatically by correlating it to human summaries. There are various variations of ROUGE such as ROUGE-n, ROUGE-L and ROUGE-SU. In ROUGE-n, a series of n-grams is elicited from the human summaries used as reference and the candidate summary. ROUGE-L used the longest common subsequence (LCS) approach i.e. the longer the LCS, more will be the similarity. The metric

ROUGE-SU makes use of bi-grams as well as uni-grams.

Results of our evaluation are shown in a given table below:

Rouge Type	Task Name	Average Recall	Average Precision	Average FScore	Number Referenced Summaries
ROUGE1	Sample 1	1.0	0.29664	0.45732	1
ROUGE1	Sample 2	1.0	0.09125	0.16841	1
ROUGE1	Sample 3	1.0	0.33504	0.50192	1

Table 1. Results of summary evaluation using ROUGE 2.0 Evaluation Toolkit

VIII. TEXT TO SPEECH SYNTHESIS

The text-to-speech synthesis [7] is the self-regulating conversion of a text into speech by transcribing the text into phonetic representation and then generates the speech waveform. A text-to-speech consists of a front-end and a back-end. The front-end performs two major operations which are text normalization and assigning phonetic transcription to each word (text-to-phoneme). The back-end part generates the speech waveform. The engine is divided into modules such as Natural Language Processing (NLP) module, Digital Signal Processing (DSP) module, text analysis and application of pronunciation rules. This can be developed using Java programming language.

There are various techniques to preform speech synthesis like Concatenative synthesis, Articulatory synthesis, Formant synthesis, Domain Specific synthesis, Unit selection synthesis, Diphone synthesis, HMM based Synthesis etc. Concatenative synthesis involves concatenation of short samples of speech recording. Articulatory synthesis makes use of articulatory parameters like human vocal tract to generate speech. Formant Synthesis is clear at high speeds. It is rule based synthesis which synthesize speech using acoustic rules. Domain-specific synthesis uses a simple approach of concatenating pre-recorded words and phrases to complete a sentence.

Unit selection synthesis makes use of segmented records stored in database to create speech.

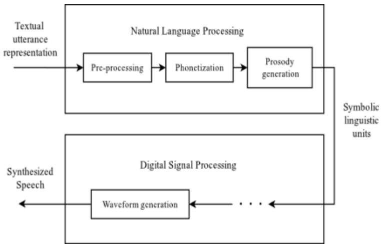
Textual utterance representation
Natural Language Processing
Pre-processing
Phonetization
Prosody generation
Symbolic linguistic units
Digital Signal Processing
Synthesized Speech
Waveform generation
. . .

Fig2. Overview of conversion of text to speech

Fig 2 shows how speech is generated from text. Natural Language Processing analyze and synthesize natural language and speech.

IX. CONCLUSION & FUTURE SCOPE

The paper proposes an approach to generate an optimized summary taking input from various mode such as image, audio and editable text. We also talked about different summarization techniques such as abstraction and extraction based. In order to generate summary we proposed modification to graph based algorithm i.e. TextRank algorithm. Besides the entire paragraph, score of the paragraph title is also taken account. Three sample articles are computed using ROUGE evaluation toolkit and the results are depicted in table 1. However there is a scope for video analysis. Since the paper discusses about taking input from multiple modes, video can be amongst them. Also, improvements can be made to make the summary algorithm more efficient and accurate. This will in turn ensure that the generated summary has its logical meaning.

REFERENCES

[1] Mrunmayee Patil, Ramesh Kagalkar, "*A Review on Conversion of Image to Text As Well As Speech Using Speech Detection and Image Segmentation*", International Journal Of Science and Research.

[2] S. Shahnawaz Ahmed, Shah Muhammed Abid Hussain and Md. Sayeed Salam, "*A Novel Substitute for the Meter Readersin a Resource Constrained Electricity Utility*" IEEE Trans. On Smart Grid, vol. 4, no. 3, Sept. 2013.

[3] 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, vol.2, Sept. 2014.

[4] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi

Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut, "*Text Summarization Techniques: A Brief Survey*", 28 Jul. 2017.

[5] Sunchit Sehgal, Badal Kumar, Maheshwar Sharma, Lakshay Rampal, Ankit Chaliya, "*A Modification to graph based approach for extraction based Automatic Text Summarization*", Institute of Electrical and Electronics Engineer.

[6] Chirantana Mallick, Ajit Kumar Das, Madhurima Dutta, Asit Kumar Das, Apurba Sarkar, "*Graph-based Text Summarization Using Modified TextRank*", Aug. 2018.

[7] Itunuoluwa Isewon, Jelili Oyelade, Olufunke Oladipupo, "*Design and Implementation of Text to Speech Conversion for Visually Impaired People*", International Journal of Applied Information System, vol.7 no.2, Apr. 2014.

[8] Kaladharan N, "*An English Text to Speech Conversion*", International Journal of Advanced Research in Computer Science and Software Engineering, vol.5, Oct. 2015.

[9] R. Aida-Zade, C. Aril, A.M. Sharifova, "*The Main Principles of Text-to-Speech Synthesis System*", International Journal of Computer and Information Engineering, vol. 7 no. 3, 2013.

[10] S. Brin and L. Page, "*The anatomy of a large-scale hypertextual Web search engine*", Computer Networks and ISDN systems, 30(1-7), 1998.

[11] Horacio Saggion and Thierry Poibeau. 2013. "*Automatic text summarization: Past, present and future*". In Multi-source, Multilingual Information Extraction and Summarization. Springer, 3–21.

[12] Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. "*Document Summarization Using Conditional Random Fields*". In IJCAI, Vol. 7. 2862–2867.

[13] Sérgio Soares, Bruno Martins, and Pavel Calado. 2011. Extracting biographical sentences from textual documents. In Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011), Lisbon, Portugal. 718–30.

[14] Karen Spärck Jones. 2007. "*Automatic summarizing: The state of the art. Information Processing & Management*" 43, 6 (2007), 1449–1481.

[15] Josef Steinberger, Massimo Poesio, Mijail A Kabadjov, and Karel Ježek. 2007. Two uses of anaphora resolution in summarization. Information Processing & Management 43, 6 (2007), 1663–1680.

[16] Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. Handbook of latent semantic analysis 427, 7 (2007), 424–440.

[17] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web. ACM, 697–706.

[18] Simone Teufel and Marc Moens. 2002. "*Summarizing scientific articles: experiments with relevance and rhetorical status*". Computational linguistics 28, 4 (2002), 409–445.

[19] Marcia A. Bush, "Speech and Text-Image Processing in Documents", Technical Report P92-000150, Xerox PARC, Palo Alto, CA, November, 1992