**RESEARCH PAPER ON**

**“Resonate- a website for converting text to speech”**

Submitted for partial fulfillment of award of the degree

**BACHELOR OF TECHNOLOGY**

**(Information Technology)**

**BY**

 Aaron Mathews (MITU18BTIT0030)

 Adia Sinu Jacob (MITU18BTIT0015)

 Gauri Wankhade (MITU18BTIT0060)

 Musharraf Sukhpurwala (MITU18BTIT0056)

**Under the guidance of**

Prof. Rohini Bhosale



**Department of Information Technology**

**MIT School of Engineering**

**MIT Art, Design and Technology University, Pune MAEER’s Rajbaug Campus, Loni-Kalbhor, Pune 412201** 2021-2022

**Abstract:**

Synthesizing speech is quite complex as it is heavily reliant on language. Meaning, the language processing section in a TTS(Text to Speech) system inherently has the largest chunk of linguistic knowledge for a particular language. The technical as well as theoretical challenges faced while building such a high-quality system can be quite daunting and hard to navigate. To ensure that the system has relevant and updated linguistic information, one must make sure it has access to the most natural and unrestricted text to ensure quality and authenticity. We will also need extensive studies to achieve the same. This paper aims to present certain issues faced while developing a TTS software in various languages and puts forth necessary research for the same. It emphasizes the variations and irregularities occuring in a language and how that poses as a barrier to improving the precision of the language processing. The first section describes our standpoint on TTS synthesis as well as the overall architecture of our TTS system. The next sections will put forth detrimental tasks in the processing stage and explains the design ideology of the modules and implementation decisions. The paper ends with conclusions and a description of the work that is currently in progress at other levels of the TTS system.

**Introduction:**

This Website is about creating a sound that represents an uploaded pdf or image. Text-to-Speech technology reads aloud digital text. It can take words on computers, smartphones, tablets and convert them into audio. Also, all kinds of text files can be read aloud, including word, pages document, online web pages. Aim to help kids who struggle with reading. Aim to help kids who struggle with reading. Text-to-Speech (TTS) is a type of assistive technology that reads digital text aloud. It’s sometimes called “read aloud” technology. TTS can take words on a computer or other digital device and convert them into audio. With a click of a button or the touch of a finger, TTS can take words on a computer or other digital device and convert them into audio.

**Literature Survey**

**Title:** Text Extraction From Image and Text to Speech Conversion

**Journal details:** Published (First Online): 13-02-2021 ISSN (Online) : 2278- 0181

**Publisher Name** : IJERT

**Limitations:** Document images are not preprocessed for noise reduction and normalization. However, global parameters, such as lower-upper baseline and slant angle are estimated .

The proposed method successfully detects the text regions in most of the images and is quite accurate in extracting the text from the detected regions. Based on the experimental analysis that we performed we found out that the proposed method can accurately detect the text regions from images which have different text sizes, styles and color. Although our approach overcomes most of the challenges faced by other algorithms, it still suffers to work on images where the text regions are very small and if the text regions are blurred.

**Title:** Multilingual Speech and Text Recognition and Translation using Image

**Journal details:** Published (First Online): 04-04-2016 ISSN (Online) : 2278- 0181

**Publisher Name** : IJERT

**Limitations:** Different nonlinear models such as ANN (Artiﬁcial neural network) and SVM (Support vector machines) can be explored with various features for developing accurate prosody models. Still more insightful research can be done on concatenation cost and target cost calculation based on syllable speciﬁc characteristics.

In this proposed system, we implemented the system for user who phasing problems of language barrier and also it user interface is also user friendly so that user can easily interact with this system .so because of this system don’t have to use dictionary for understanding the meaning of word, so it automatically reduce the user task for understanding the languages for communication.

**Title:** Development of syllable-based text to speech synthesis system

**Journal details:** Received: 6 January 2011 / Accepted: 31 May 2011 / Published online: 16 June 2011

**Limitations**: suffers to work on images where the text regions are very small and if the text regions are blurry.

less effective when the text is too small and if the text region is not clearly visible or the color of the text is not visible clearly.

In this work, a prototype Bengali TTS using syllables as the basic unit was developed using the festival framework. Text corpus is collected from various domains. Optimal text selection algorithm is used to reduce redundancy in the text corpus. Different nonlinear models such as ANN (Artificial neural network) and SVM (Support vector machines) can be explored with various features for developing accurate prosody models.

**Title:** Optical Character Recognition for Cursive Handwriting

**Journal details**: IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 6, June 2002

**Limitations:** Audio not loud and clear. User interface difficult.

This study is the development of a powerful segmentation algorithm. This is one of the studies, which fine tunes and completes the approaches presented in the previous studies to a certain extent, thus providing some improvements in the algorithms and recognition rates.

**Title:** Designing, development and implementation of Text to

Speech

**Journal details**: Text to speech, Concatenative Synthesis, Gujarati

consonants and vowels, TTS IDE., September 2005

Algorithm and implementation of Gujarati TTS system using phoneme concatenative methodology was developed by researchers. This project focuses primarily on the process of creating a voice for a concatenative Text-To-Speech system, or altering the TTS system's own standard output voice to sound more like the target voice.

**Existing Systems:**

Image to Text Converter

•We present an online OCR (Optical Character Recognition) service to extract text from images. Upload a photo to our image to text converter, click on submit and get your text file instantly.

•Convert your images to text. Extract text from images, photos, and other pictures. This free OCR converter allows you to grab text from images and convert it to a plain text TXT file. We currently support the following image to text conversions: JPG to text, PNG to text, TIFF to text, SVG to text, BMP to text, WEBP to text, and many more!

•Drag and drop your files, or type, paste, and edit text here. Natural Reader isa professional text to speech program that converts any written text into spoken words. The paid versions of Natural Reader have many more features.

**Proposed System:-**



**Architecture diagram:-**



**Algorithm used:**

The website works solely on python framework, for the project python 3.6.7 is used. As the project idea describes, an uploaded image is converted into audible mp3 format. So, the process is divided into six parts:

1. Getting the image from the user on the website.

2. Reading the image.

3. Process Image.

4. Image to Text.

5. Text to speech.

6. Return audio file

**Results:**

User will sign in to our portal and will be able to go to the convert page where the user can see the upload button. By clicking on the upload button, the user can upload the image file that is to be converted.

After the user uploads the image file (JPG/JPEG, PNG) or a document (PDF/docx, doc) he wants to convert after clicking the upload button the audio file will be automatically downloaded so that the user can listen to the file offline too.

Handwritten files can also be uploaded, converted and downloaded.







**Conclusion and future work:**

Our site gives the client a sound record of the text/image/document. It could be a web-based stage which permits the client to transfer any picture with content and changes over them into any sound arrangement. Offices of downloading the m record, playing it on the browser or getting it sent are there. This way the client has numerous ways to get a sound record. Text to speech synthesis is an aspect of computer technology that is growing at a very great pace and it is playing a very crucial role in the way that the user interacts with the virtual system and interfaces across a variety of platforms.

Identification of various operations and processes that are related to text-to-speech synthesis has been made. With the use of the code and the library mentioned above in python we have achieved the normal text to speech conversion in the specified language. In future there is a huge scope in this field. The user can train a dataset for a different voice using speech recognition and use it as voice for output of the speech that we receive for the input text that we give.

Currently the program is limited to the English language; however, users might upload an image in some other language where the program fails. TesseractOCR is capable of recognizing various languages. Hence for future scope this might be a goal to implement various languages in the project.

**Future Scope:-**

**Accessibility:** Text-to-speech solutions provide improved digital accessibility to populations with learning and speech disabilities, visual impairments, and low literacy across devices and platforms. Audio enabled website and Augmented and Alternative Communication (AAC) devices and other communication devices used by those with a speech impairment.

**Automotive:** Can be effectively used in navigation systems and GPS, Outbound correspondences among showrooms and clients for things like arrangement affirmations, planned assistance updates, and advancement and deals updates can without much of a stretch be computerized utilizing one of the resonate applications. Government websites can be made read aloud hence they can reach to each and every person. It can be used for emergency alerts and speech enabled tax visa fillings

**Health:** Can be effectively used in health monitoring, medical devices, dial in pharmacy and appointment reminders.

**Bibliography:**

[1]Dutoit, T., 1993. High quality text-to-speech synthesis of the French language. Doctoral dissertation, Faculté Polytechnique de Mons.

[2] Suendermann, D., Höge, H., and Black, A., 2010. Challenges in Speech Synthesis. Chen, F., Jokinen, K., (eds.), Speech Technology, Springer Science + Business Media LLC.

[3] Allen, J., Hunnicutt, M. S., Klatt D., 1987. From Text to Speech: The MITalk system. Cambridge University Press.

[4] Rubin, P., Baer, T., and Mermelstein, P., 1981. An articulatory synthesizer for perceptual research. Journal of the Acoustical Society of America 70: 321–328. [5] van Santen, J.P.H., Sproat, R. W., Olive, J.P., and Hirschberg, J., 1997. Progress in Speech Synthesis. Springer.

[6] van Santen, J.P.H., 1994. Assignment of segmental duration in text-to-speech synthesis. Computer Speech & Language, Volume 8, Issue 2, Pages 95–128

[7] Wasala, A., Weerasinghe R. , and Gamage, K., 2006, Sinhala Grapheme-to-Phoneme Conversion and Rules for Schwaepenthesis. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, Australia, pp. 890-897.

[8] Lamel, L.F., Gauvain, J.L., Prouts, B., Bouhier, C., and Boesch, R., 1993. Generation and Synthesis of Broadcast Messages, Proceedings ESCA-NATO Workshop and Applications of Speech Technology.