iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44060

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Resonate: Website on Text to Speech

Gauri Wankhade[1], Musharraf Sukhpurwala[2], Adia Sinu[3], Aaron Mathews[4], Prof. Rohini Bhosale[5]
[1, 2, 3, 4]*Students, Dept of IT, MIT School of Engineering, Pune, Maharashtra, India*
[5]*Professor, Dept of IT, MIT School of Engineering, Pune, Maharashtra, India*

Abstract: ***Synthesizing speech is quite complex as it is heavily reliant on language. Meaning, the language processing section in a TTS 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. At the heart of this software engine lies an OCR Engine (Optical Character Recognizer) which inherits crucial morphological operations required for image conditioning & transformation, accompanied with python libraries used for character classification. Further the processed textual data is transformed into speech signals using various Text-to-Speech synthesis techniques...***
Keywords: Speech, TTS(Text-to-Speech), OCR, Transformation, Text-to-Speech.

I. INTRODUCTION

Resonate 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.

Also, all kinds of text files can be read aloud, including word, pages document, online web pages. TTS is an 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.

Resonate implements a python-based system that can translate text written into English language as per the user's requirement with correct grammatical sequencing and anaphorical resolution, and implement an optical character recognizer with a reasonably high accuracy that can identify each and every optical/lingual/mathematical character imprinted on the image or imagery module under consideration.

Text-to-speech synthesis (TTS) is the automatic conversion of a text into speech that sounds as if it were read by a native speaker of the language. TTS is a technology that allows a computer to speak to you. The text is fed into the TTS system, which then uses a computer algorithm called the TTS engine to analyse, pre-process, and synthesis speech using mathematical models. The output of the TTS engine is usually sound data in an audio format.

Speech sound is finally generated with the low-degree synthesizer with the aid of using the information from high-degree one.

II. LITERATURE REVIEW

The proposed technique efficiently detects the textual content areas in maximum of the image and is pretty correct in extracting the textual content from the detected areas. Based at the experimental evaluation that we executed we observed out that the proposed technique can appropriately stumble on the textual content areas from image that have one of a kind textual content sizes, patterns and colour. Although our approach overcomes maximum of the demanding situations confronted via way of means of different algorithms, it nonetheless suffers to paintings on image wherein the textual content areas are very small and if the textual content areas are blur. [1]

In this paper, a whole laptop software is offered that may convert Bangla PDF to Bangla Speech. According to the proposed technique, photographs are extracted from PDF after which after processing the photographs, they're despatched to OCR engine to extract textual content. Extracted textual content are then normalized and despatched to textual content to speech (TTS) engine to generate speech. Image processing is a key aspect of the evolved software because it will increase the performance of OCR engine to a amazing extent. We suggest a unique threshold choice approach this is capable of hit upon sort of noise withinside the extracted photograph and pick out threshold for that reason for binary transformation. Thus, it solves the trouble of choosing suitable threshold of various photographs and it will increase the general accuracy and performance of the software.

Another characteristic that has advanced the overall performance of delivered laptop software is textual content normalization. Normalization of the extracted textual content from the OCR engine makes the textual content extra correct to pronounce with the aid of using the TTS engine relying at the context. Finally, we gift experimental effects that display 80.804 accuracy on textual content extraction from the PDF document and 3. ninety-two score (out of 5) at the generated speech with the aid of using human evaluation. [2].

In TTS-primarily based totally audiobook production, multi-function dubbing and emotional expressions can notably enhance the naturalness of audiobooks. However, it calls for guide annotation of authentic novels with express speaker and emotion tags in sentence stage, that is extraordinarily time-eating and costly. In this paper, we suggest a chapter-clever know-how machine for Chinese novels, to are expecting speaker and emotion tags routinely primarily based totally at the chapter-stage context. Compared with baselines of every component, our fashions attain better performance. Audiobooks produced with the aid of using our proposed machine together with a multi-speaker emotional TTS machine, are proved to attain similar great rating to audiobooks made with the aid of using man or woman producers. [3].

This paper includes Desktop and Mobile packages with TTS verbal exchange analysis. Normally, human beings are aware about the Desktop and Mobile packages. This paper includes all functionalities of cell packages for that researcher have indexed many packages and researchers have taken a number of them on bases of the recognition of cell packages, and from that created chart and characteristic extraction of all cell packages for higher understanding. This paper additionally includes the statistics of computing device packages with TTS functionalities, and researchers have indexed the all-computing device packages which can be presenting the TTS functionalities [4].

III. PROPOSED METHODOLOGY

A. Proposed Method

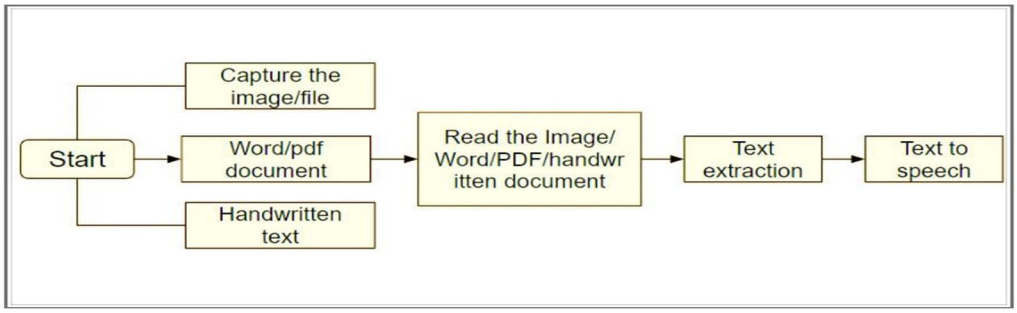


Figure 1. Block Diagram of the modules

The basic block diagram (refer to figure 1) project includes a method for capturing input text via image, pdf, or word file. The read the image block comes next, where the manipulated text is read and ready to be converted into speech using unit selection synthesis. The text extraction module is manipulated in the capture module, which changes the voice, rate, and volume. We can create an image-to-speech module by combining the modules listed above.

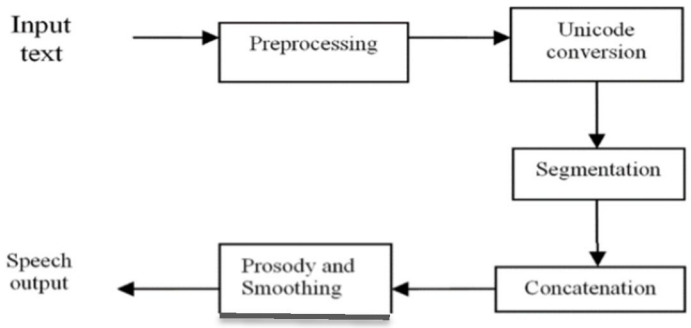


Figure 2. Text-to-Speech Problem Formulation.

Figure 2 depicts the flow of the text-to-speech module, which is described in detail below: Beginning with the method of pre-processing the input text, text to speech conversion can be accomplished.*B. Algorithm*
The internet site works completely on python framework, for the challenge python 3.6.7 is used. As the venture concept describes that an uploaded photograph is transformed into audible mp3 format. So, the system of is split into six parts:

B. Algorithm

The internet site works completely on python framework, for the challenge python 3.6.7 is used. As the venture concept describes that an uploaded photograph is transformed into audible mp3 format. So, the system of is split into six parts:

1) Getting the image from user in website.
2) Reading the image.
3) Process Image
4) Image to Text.
5) Text to speech.
6) Return audio file

a) *Getting the Image from User in Web App:* A web-based platform is provided for user to interact with Image to Speech API. Users need to upload an image in the web form. POST' approach is used to ship the photograph to python script.
b) *Reading the Image:* Once the acquired via way of means of python script Open-CV's cv2 elegance is used to examine the photo and convert it into gray scale photo of 1's and 0's.
c) *Process Image:* The array of photo is similarly dilated and eroded to make the photo noise unfastened, and the pixels of photo are increased, and it's far transformed into grey scale in order that it may be equipped to byskip to OCR Engine.
d) *Image to Text:* The noise unfastened clean photo is surpassed to TesseractOCR engine. The OCR engine converts the typed facts withnside the photo to string format. The string's accuracy relies upon at the readability of the picture, but in a few instances, there are errors in conversion of punctuation symbols like (",'!).

C. Working

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.

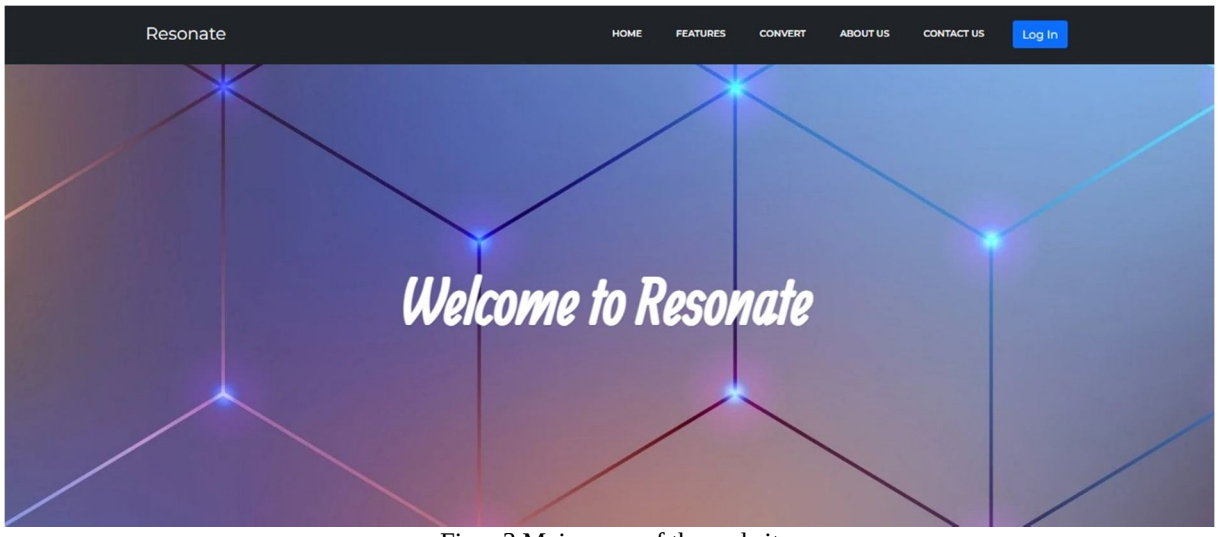


Fig – 3 Main page of the website

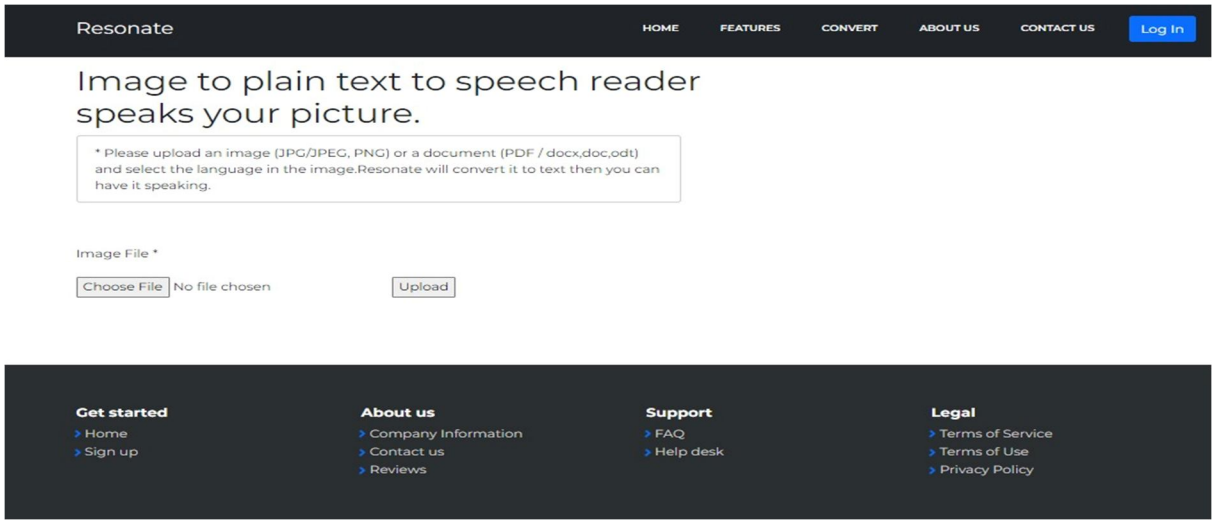


Fig - 4: The upload and convert page

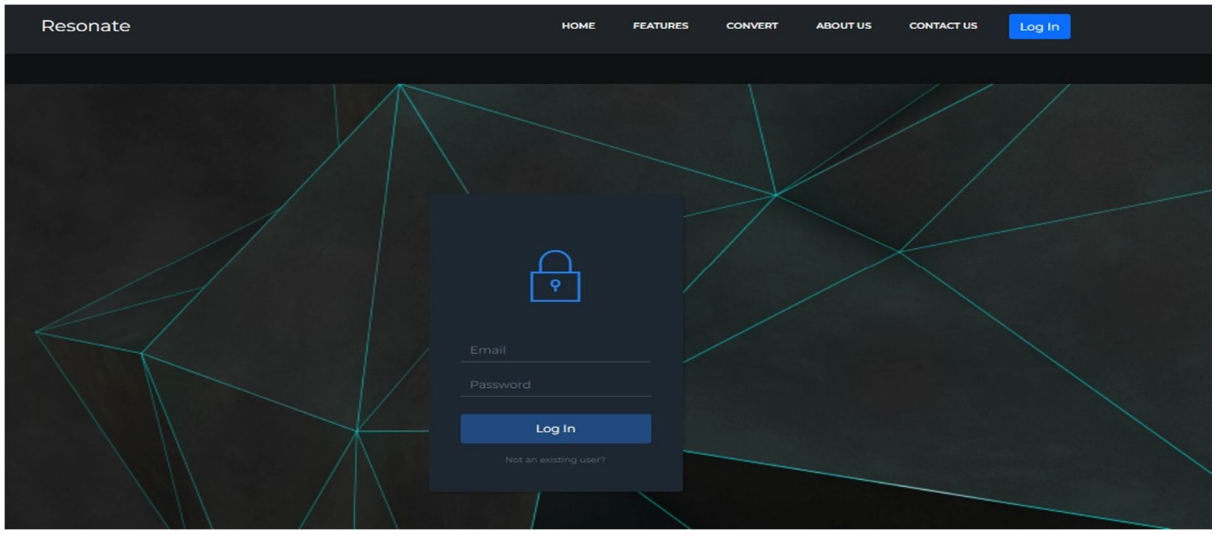


Fig - 5: Login Page

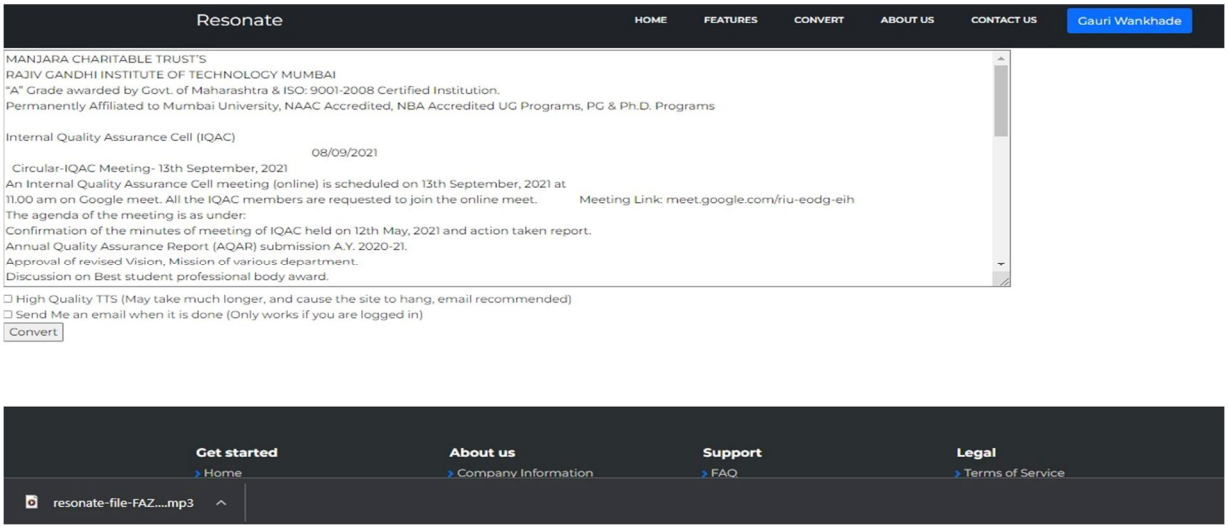


Fig - 6: The file is read and displayed

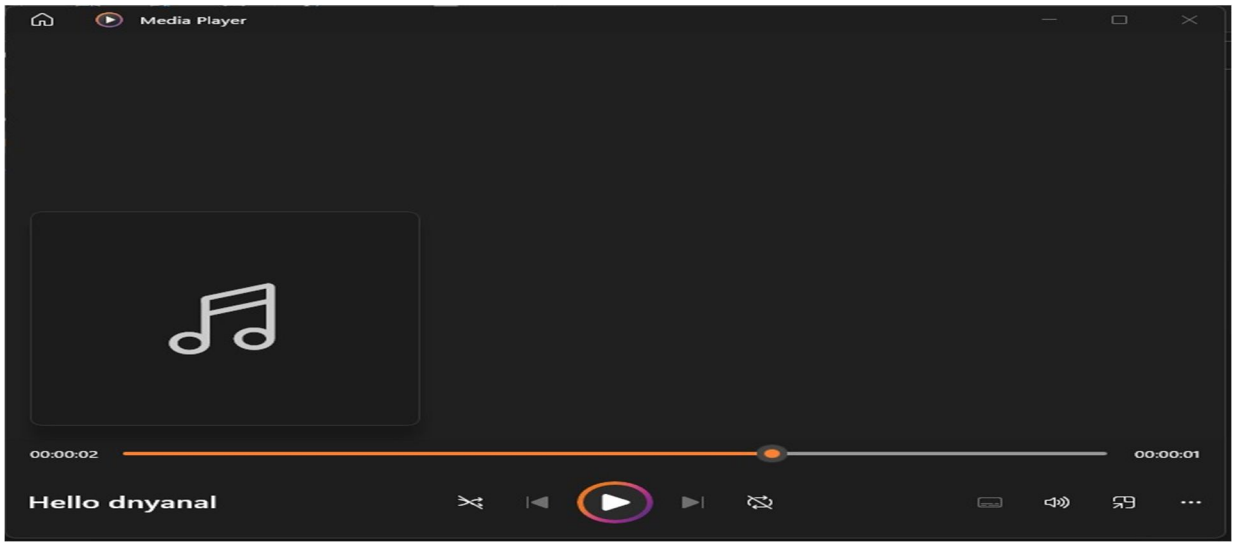


Fig - 7: The downloaded audio file can play offline also

IV. CONCLUSIONS

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. Tesseract is capable of recognizing various languages. Hence for future scope this might be a goal to implement various languages in the project.

A. 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 internet site and Augmented and Alternative Communication (AAC) gadgets and different conversation gadgets utilized by people with a speech impairment.

Automotive: Can be efficaciously utilized in navigation structures and GPS, Outbound correspondences amongst showrooms and customers for such things as association affirmations, deliberate help updates, and development and offers updates can without much of a stretch be computerized utilizing one of the resonate applications.

Government web sites may be made study aloud therefore they are able to attain to every and each person. It may be used for emergency indicators and speech enabled tax visa fillings

Health: Can be successfully utilized in fitness monitoring, scientific devices, dial in pharmacy and appointment reminders.

REFERENCES

[1] Text Extraction From Image and Text to Speech Conversion Journal details: Published (First Online): 13-02-2021 ISSN (Online) : 2278- 0181 Publisher Name : IJERT

[2] Bangla PDF Speaker : A Complete Computer Application to Convert Bangla PDF to Speech Published in: 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI) Date of Conference: 08-09 July 2021 Publisher: IEEE Conference Location: Rajshahi, Bangladesh.

[3] A Chapter-Wise Understanding System for Text-To-Speech in Chinese Novels Published in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Date of Conference: 06-11 June 2021 Publisher: IEEE Conference Location: Toronto, ON, Canada.

[4] Afsana Kargathara, Krishna Vaidya & C. K. Kumbharana ," Analysing Desktop and Mobile Application for Text to Speech Conversation", Conference paper

published on 02 October 2020.

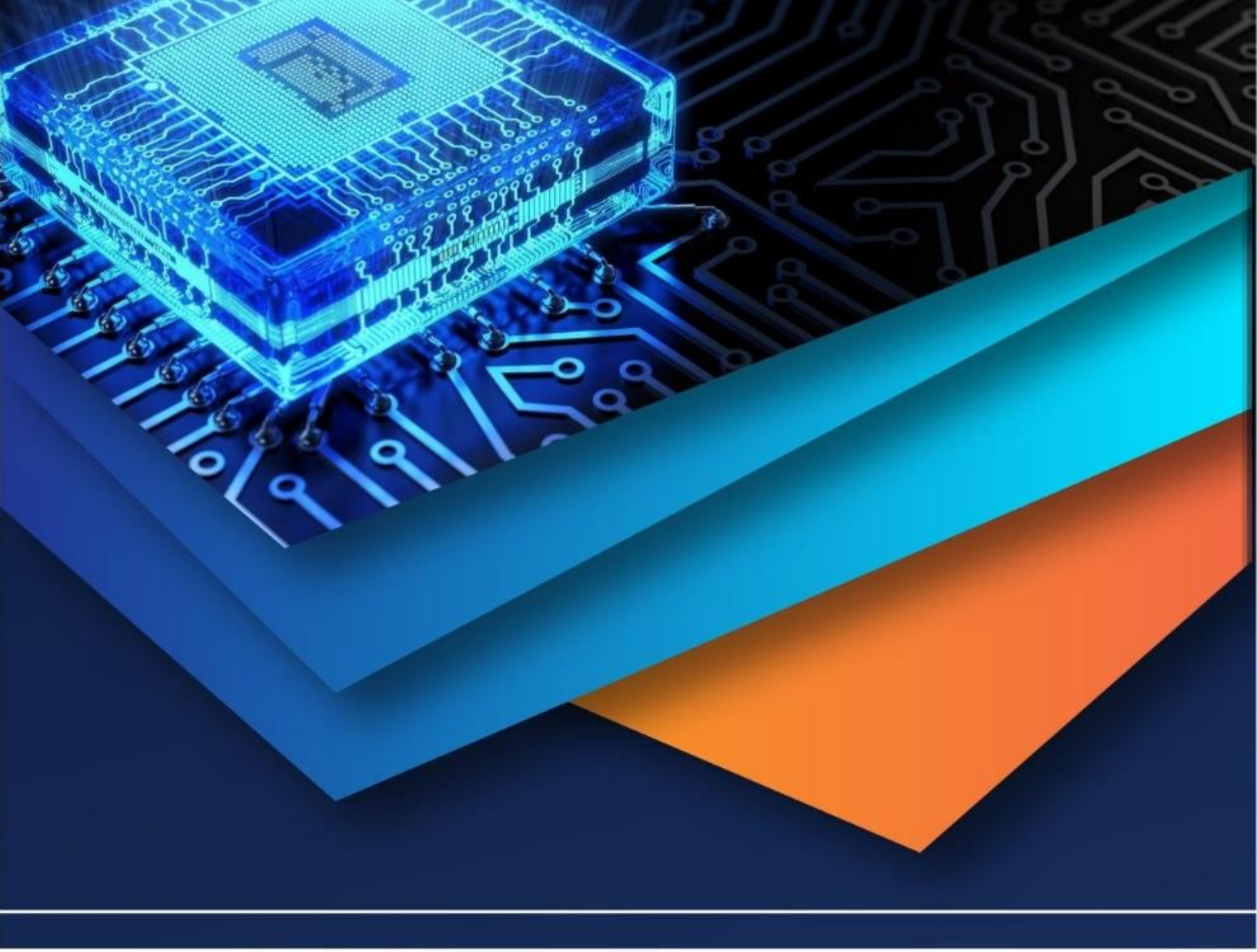
[5] Shen, J., Jia, Y., Chrzanowski, M., Zhang, Y., Elias, I., Zen, H., and Wu, Y. Non-Attentive Tacotron: Robust and Controllable Neural TTS Synthesis Including Unsupervised Duration Modelling. ArXiv, abs/2010.04301, 2020.

[6] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-Based Generative Modelling through Stochastic Differential Equations. In International Conference on Learning Representations, 2021.

[7] Prenger, R., Valle, R., and Catanzaro, B. Waveglow: A Flow-based Generative Network for Speech Synthesis. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3617–3621. IEEE, May 2019.

[8] Ren, Y., Ruan, Y., Tan, X., Qin, T., et al. FastSpeech: Fast, Robust and Controllable Text to Speech. In Advances in Neural Information Processing Systems 32, pp. 3171– 3180. Curran Associates, Inc., 2019.

[9] Yamamoto, R., Song, E., and Kim, J.-M. Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6199–6203, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY