I.HEAR.YOU: A Web-Based Image-To-Speech Converter Application Using Optical Character Recognition and Speech Synthesis

Mulya Fajar Ningsih Alwi
Faculty of Computing
President University
Bekasi, Indonesia mulya.alwi@student.president.ac.id

***Abstract*--- Image-to-Speech technology is an assistive technology that enables visually impaired individuals to "see" images through audio descriptions. It utilizes Optical Character Recognition (OCR) and Text-to-Speech (TTS) Synthesizer, employing Natural Language Processing (NLP) and Digital Signal Processing (DSP) to analyze image contents and generate spoken descriptions in various languages. This technology is valuable for navigating surroundings, comprehending visual content, and enhancing the quality of life for visually impaired individuals. The final project focuses on converting image text into audio, supporting multiple languages, and facilitating accessibility for illiterate individuals. The accuracy of OCR performance was tested, resulting in high precision, recall, and an average accuracy rate of 0.976, demonstrating its effectiveness. The experiments provide insights for improving real-world application performance.**

Keywords: Image-to-Speech, Optical Character Recognition, Speech Synthesis, Natural Language Processing, Digital Signal Processing.

I. INTRODUCTION

Technology plays a crucial role in addressing various human challenges, even in smaller-scale tasks like household chores. In Indonesia, image-to-speech technology has emerged as a valuable tool with advancements in digital signal processing and machine learning. These improvements have enhanced accuracy and efficiency, benefiting applications such as image recognition, augmented reality, and accessibility tools. However, the complexity of visual information poses a challenge that requires natural language processing and machine learning algorithms to generate accurate audio descriptions. Despite these challenges, image-to-speech technology has the potential to revolutionize digital content consumption and interaction. Indonesia faces significant literacy challenges, ranking low in literacy rates compared to other countries. To tackle this issue, a proposed project aims to develop a technology solution that aids in literacy comprehension and assists visually impaired individuals, particularly those who are illiterate. Utilizing deep learning, machine learning, and artificial intelligence techniques, the project seeks to recognize image-based sentences and convert them into audio files, enabling individuals to access information effectively. The goal is to leverage technology to bridge the literacy gap and empower visually impaired individuals to access information through image-to-speech technology.

II. LITERATURE REVIEW

A. Artificial Intelligence (AI)

Artificial intelligence (AI) is a rapidly evolving field focused on creating human-like machines. It includes machine learning, deep learning, natural language processing, and computer vision. AI revolutionizes industries with applications like self-driving cars and smart assistants. Machine learning improves machine performance, while natural language processing enables language comprehension. Computer vision helps machines interpret visuals. Despite its potential, ethical concerns like job displacement and bias must be addressed in AI's implementation.

B. Image Recognition

Image recognition, also called image classification, uses AI and machine learning to identify and categorize objects in digital images or videos. The process involves collecting labelled images, pre-processing them, extracting features, training the algorithm, and testing its accuracy. Techniques include object and facial recognition, scene recognition, OCR, and gesture recognition, with applications in multiple fields. Image recognition aims to make machines understand visual data like humans, offering automation and efficiency. However, ethical concerns about privacy and surveillance need to be addressed.

C. Optical Character Recognition (OCR)

Optical Character Recognition (OCR) converts printed or handwritten text and text images into digital data. The process involves scanning, pre-processing, character recognition, and post-processing. Scanning captures the document, pre-processing enhances it, and OCR algorithms recognize characters, converting them to ASCII format. Post-processing retains structure, creates searchable PDFs, and may involve NLP for verification. OCR achieves high accuracy with familiar text but may have errors with handwriting or unfamiliar documents, requiring monitoring and correction of output.

D. Speech Synthesis

Speech Synthesis, or Text-to-Speech (TTS), is computer-generated human-like speech. It converts written text into spoken words and is used in voice-enabled services and assistive technology for visually impaired individuals. The process involves analyzing text, generating phonemes, determining prosody, and converting it into audio. Advancements in technology have improved the naturalness of synthesized speech.

E. Natural Language Processing (NLP)

Natural Language Processing (NLP) combines linguistics, computer science, and AI to enable computers to process and analyze human language. It aims to understand and extract information from large language datasets, including speech recognition and generation. In Text-to-Speech (TTS) systems, NLP generates phonetic representation, analyzing text structure, context, and meaning through steps like segmentation, tokenization, parsing, and tagging. Techniques like normalization and language modeling improve synthesized speech authenticity. NLP is crucial in TTS to ensure natural-sounding speech that effectively conveys text meaning.

F. Digital Signal Processing (DSP)

Digital Signal Processing (DSP) is crucial in Text-to-Speech (TTS) systems, converting symbolic information from NLP into understandable speech. Using mathematical algorithms, DSP modifies and analyzes digital signals for natural and clear synthesized speech. Rule-based synthesis adjusts parameters dynamically, while concatenative synthesis strings together speech segments for a natural output. Filtering and noise reduction enhance speech quality, accurately representing sounds, timing, pitch, and duration. DSP ensures high-quality, expressive TTS that effectively conveys meaning and emotion from input text.

III. SYSTEM ANALYSIS

A. System Overview

The project aims to develop a web-based image-to-speech converter application called I.HEAR.YOU. It utilizes AI technologies such as OCR and Speech Synthesis, incorporating NLP and DSP components. The application allows users to upload images in JPG, JPEG, or PNG formats containing text. Using OCR, the application recognizes the image's content and converts it into text. The detected text is then synthesized into an audio file, spoken in various languages supported by the application, utilizing Speech Synthesis with NLP and DSP. After uploading an image, users are redirected to an image-decoded page displaying the uploaded image, the detected text's language, and the text itself, along with autoplaying the audio. The application also offers image translation into supported languages. Users can select a translation language, initiate the translation process, and view the translated text while the audio is played using the correct language's accent. Additionally, users have the option to download the audio file of the translated text. The primary goal of this user-friendly application is to assist visually impaired individuals and those with limited literacy skills in accessing image information conveniently through audio files. The project utilizes Python, the Flask framework, and various Python libraries like OpenCV, Tesseract, gTTS, googletrans, Flask-WTF, and Flask-Dropzone to develop the application.

B. Functional Analysis

There are some features and functionality in this project that are aimed to be fully implemented in order to maximize the goals of the project and give the user an experience that they will enjoy. The descriptions of these functionalities will be shown in the list below.

1) The application allows users to access the home page.
2) The application allows users to access the about page.
3) The application allows users to access the upload page.
4) The application allows users to upload an image to the provided Dropbox.
5) The application can detect the content of the uploaded image.
6) The application can autoplay the audio of the detected text in the uploaded image with the correct language accent.
7) The application can display the uploaded image, the language of the detected text, and the detected text itself.
8) The application allows users to translate the detected text in the image into various supported languages.
9) The application can display the translated text.
10) The application can autoplay the audio of the translated text with the correct language accent.
11) The application allows users to download the audio file of the translated text.

C. Use Case Diagram

A use case diagram is used to describe the flow of process, interaction, and dependencies between the user and the system. By creating a use case diagram, the functionalities, and requirements of the project could be explained easily therefore helping the development process of the project. The figure below illustrates the utilization scenario diagram for this web application project.

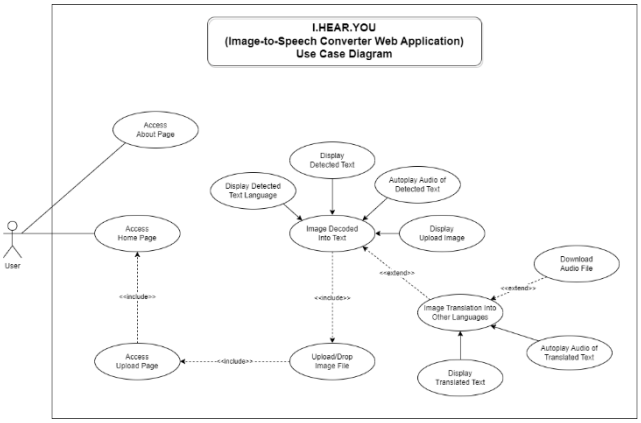

Fig. 1. Application Use Case Diagram

D. Swim Lane Diagram

A Swimlane diagram is a type of chart that shows how a process works and who is involved in it. The figures below show the swim lane diagram for the application workflow process.

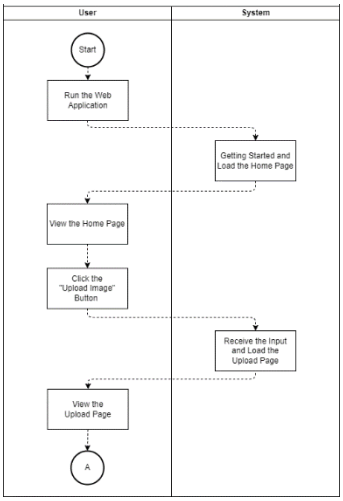

Fig. 2. Uploading Image Swim Lane Diagram

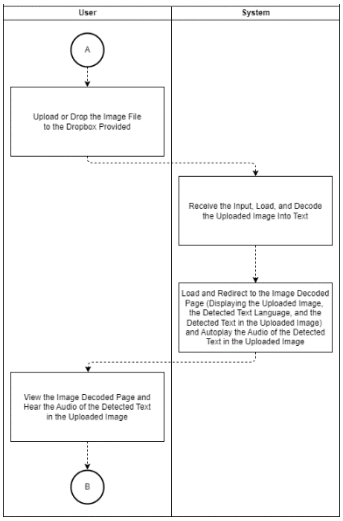

Fig. 3. Image Detection Swim Lane Diagram

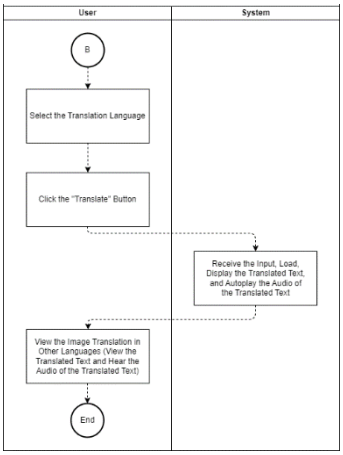

Fig. 4. Image Translation Swim Lane Diagram

IV. METHODOLOGY

Below is a figure of the application methodology, which involves two techniques, such as Optical Character Recognition (OCR) and Speech Synthesis. OCR algorithms include image acquisition, pre-processing, text detection, text recognition, and post-processing to extract the image's detected text. Speech Synthesis employs Natural Language Processing (NLP) and Digital Signal Processing (DSP), which DSP also involves three steps, such as word-to-phoneme conversion, phoneme synthesis, and waveform generation to convert the detected text into audible speech.

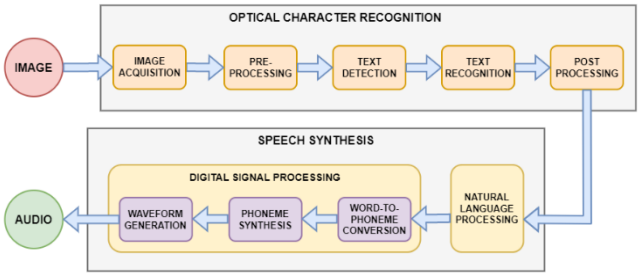

Fig. 5. Application Methodology

A. Optical Character Recognition (OCR)

The web application utilizes Optical Character Recognition (OCR) with the Tesseract OCR engine and Python pytesseract library to extract text from uploaded images. The OCR algorithm involves image acquisition, pre-processing, text detection, text recognition, and post-processing to convert detected text regions into machine-readable text for analysis.

1) **Image acquisition** is the process of obtaining an image containing the text to be recognized from sources like scans, photos, or digital images.
2) **Pre-processing** is used to enhance image quality for accurate text recognition. Techniques include resizing, noise removal, contrast enhancement, and binarization (converting to black and white).
3) **Text detection** analyzes the pre-processed image to identify text regions. Techniques like edge detection, connected component analysis, or machine learning are used to identify areas with high text likelihood, such as lines or characters.
4) **Text recognition** is the process of converting the identified text regions into editable, searchable, and machine-readable text. It involves analyzing the visual features of characters using machine learning or pattern recognition to recognize and classify them.
5) **Post-processing** is used to refine recognized text by correcting errors, improving formatting, handling special characters, and eliminating noise or artifacts introduced during OCR. It may involve language-specific processing, spell-checking, and grammar rule application.

B. Speech Synthesis

The Speech Synthesis technique converts text into synthesized speech using the Python gTTS library. It involves Natural Language Processing (NLP) and Digital Signal Processing (DSP), which the DSP also includes three steps, such as word-to-phoneme conversion, phoneme synthesis, and waveform generation. The resulting speech is saved as an audio file for playback.

1) **Natural Language Processing (NLP)** analyzes and processes input text before converting it into synthesized speech. It includes tasks like tokenization, formatting, language detection, etc. NLP aids in text translation for multilingual support, ensuring accurate and language-specific speech synthesis.

2) **Digital Signal Processing (DSP)** focuses on manipulating and processing audio signals to generate high-quality and natural-sounding speech. DSP involves three steps to produce the synthesized speech such as word-to-phoneme conversion, phoneme synthesis, and waveform generation.

    a) **Word-to-Phoneme Conversion** is the process of mapping words to their corresponding phonemes, the basic sound units in speech. It analyzes linguistic rules and context to determine the pronunciation, often utilizing a database or rule-based system for accurate representation.

    b) **Phoneme Synthesis** generates speech sounds based on phonetic representations obtained from Word-to-Phoneme Conversion. DSP algorithms model phoneme characteristics, manipulating parameters for smooth and natural transitions, resulting in intelligible speech.

    c) **Waveform Generation** is the process of converting the synthesized phonemes into an audio waveform that represents the speech. DSP techniques create high-quality waveforms resembling natural speech for playback through speakers or audio devices.

V. RESULT AND DISCUSSION

A. User Interface Development

User interface development is the real implementation of the user interface wireframe design by developing the visual design and interactive elements or components of the web application that enable users to interact with the application. The user interface (UI) of this project was developed with the help of the Python Flask Framework in order to develop the web application in a structured way. There are four main UIs that were successfully developed, such as the Home Page UI, About Page UI, Upload Page UI, and Image Decoded UI.

Below are the figures for UI design implementation along with a detailed explanation or description of how to use or interact with the web application UIs.

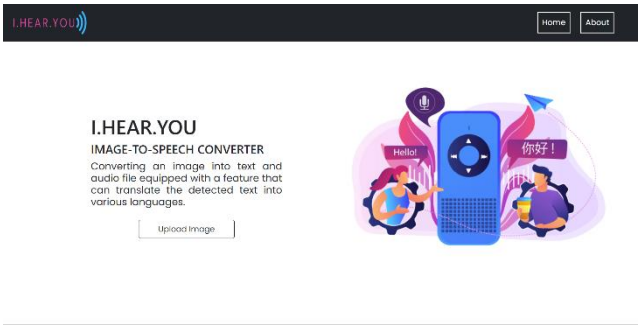

Fig. 6. Home Page User Interface

The Home Page becomes the first page that the user will access when the user runs the web application. This page becomes the main page of this web application and is used to redirect the user to other pages in the application. All the web application pages consist of the navigation bar used as navigation to another page in the application, which helps the user access the about page and mostly helps in navigating the user back to the Home Page in order to gain access to the Upload Page that is used to retry the application's feature by uploading other images without rerunning the application to access the Home Page.

Fig. 7. About Page User Interface

The About Page can be accessed if the user clicks the "About" button in the navigation bar. The About Page provides a detailed description of the web application along with its illustration image in the form of an image slideshow right next to it and also information about the application's copyright, which includes the developer's name and year of creation.

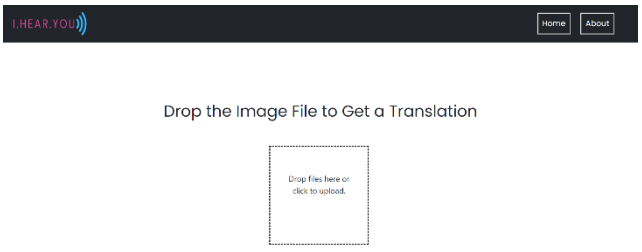

Fig. 8. Upload Page User Interface

The Upload Page can be accessed if the user clicks the “Upload Image” button on the Home Page. The Upload Page provides an image file Dropbox that the user uses to drop or upload an image file to be processed by the application. The application only allows the image format file to be uploaded and only allows one image to be uploaded at a time.

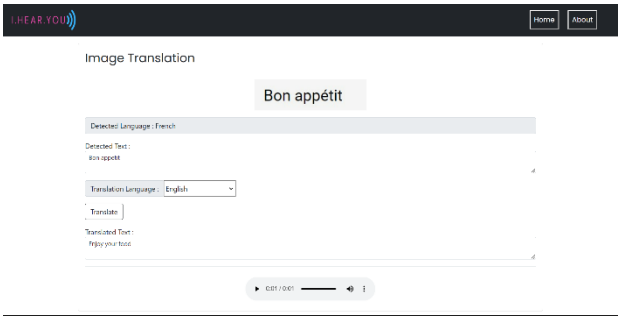

Fig. 9. Image Decoded Page User Interface

The Image Decoded Page can be accessed if the user has uploaded the image file on the Upload Page, and the uploaded image is decoded into text by the application. If the image upload and decoding are successful, then the application will redirect the user to the Image Decoded Page and autoplay the audio of the detected text in the uploaded image. The Image Decoded Page contains all the details of the uploaded image, such as the display of the uploaded image, the detected text language, and the detected text. The application also offers the feature of translation for the uploaded image into other languages supported by the application. The user can do the translation for their uploaded image on the Image Decoded Page by selecting the translation language that is provided in the selection box. After the user selects their desired translation language, they need to click the "Translate" button. The application will process the user input, then display the translated text, and autoplay the audio of the translated text with the audio controller if the translation process is successful.

B. Application Performance Testing

Application performance testing involves evaluating and measuring the performance and responsiveness of the application under various conditions. It aims to ensure that the app meets the expected performance standards and functions optimally when handling image processing, OCR, speech synthesis, and other tasks. Below is the result of the application testing in various image upload conditions and the result of the application performance calculation of accuracy, precision, and recall rate.

Tab. 1. Image Detection Testing Evaluation Result

Image Condition	Evaluation Result
Image contains a handwritten text	Detected as expected (shown in Figure 10)
Image contains a piece of text from a newspaper	Detected as expected (shown in Figure 11)
Image of the road sign contains text	Detected as expected (shown in Figure 12)
Image contains Arabic language text in low light	Detected as expected (shown in Figure 13)
Image contains ultra-thin text in the Hindi language	Detected as expected (shown in Figure 14)

Fig. 10. Handwritten Text Image Detection Testing Result

Fig. 11. Newspaper Text Image Detection Testing Result

Fig. 12. Road Sign Text Image Detection Testing Result

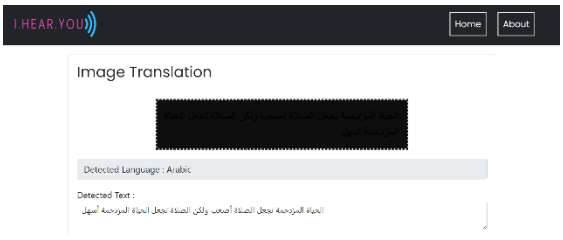

Fig. 13. Arabic Language Low Light Text Image Detection Testing Result

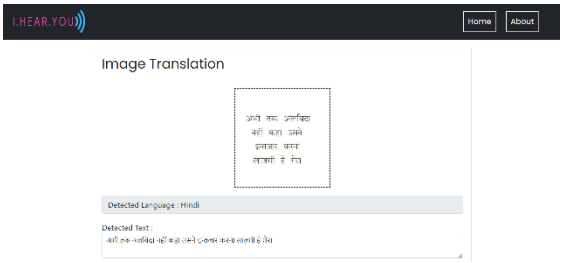

Fig. 14. Hindi Language Ultra-thin Text Image Detection Testing Result

The performance accuracy testing calculation focused on the Optical Character Recognition (OCR) technique's performance as a main feature of the web application. The performance accuracy testing calculated the accuracy rate, precision, and recall of the image detection feature in recognizing the characters contained in the uploaded image, along with the graph of the precision-recall curve. The accuracy, precision, and recall value are calculated using two samples, such as the list of words in a detected text by the OCR engine and the list of words in a ground-truth text inputted by the user, in order to calculate the number of true positives, false positives, and false negatives between the two samples. The image conditions used are an image using an English language image and a non-English language in order to know the performance accuracy of the web application in recognizing and extracting text from the given conditions. The figures below show the evaluation results of the application performance accuracy testing.

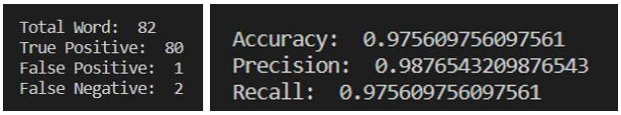

Fig. 15. English Language Text Image Accuracy Testing Result

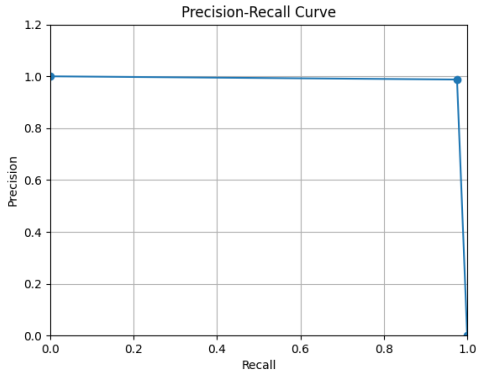

Fig. 16. Precision-Recall Curve of English Language Image Testing

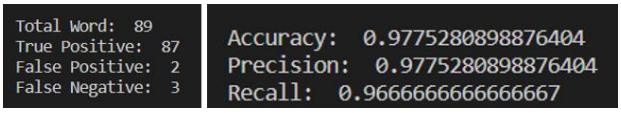

Fig. 17. Korean Language Text Image Accuracy Testing Result

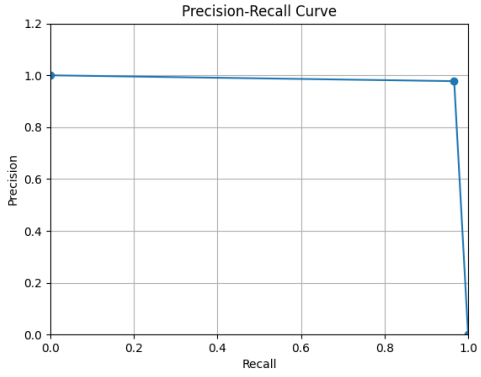

Fig. 18. Precision-Recall Curve of Korean Language Image Testing

VI. CONCLUSION AND FUTURE WORKS

In conclusion, the use of Optical Character Recognition (OCR) and Speech Synthesis technology works as intended in the web application and gives the desired output from the beginning to the end. From the experiments focused on calculating the accuracy of OCR performance and through extensive testing and analysis, the web application achieved an average accuracy rate of 0.976, an average precision of 0.982, and an average recall of 0.970, showcasing its effectiveness in accurately converting images to text. The experiments highlight the successful implementation of OCR techniques and provide valuable insights for improving the web application's performance in real-world scenarios. Hopefully, the user of this web application can get an unforgettable and enjoyable experience when using the web application by utilizing all features that are available in the web application, especially in enabling visually impaired individuals or illiterate people to access information that is contained in the image form around them in the easiest way possible using this user-friendly web application.

The web application works as intended but there are some aspects that can be improved to make the web application even better such as improving the web application into a mobile-based application that can be used online, allowing to upload and process multiple images, adding a feature that allows image detection by taking images from the camera, adding more languages to be used in a translation feature, increasing the accuracy and the processing speed of the OCR engine in recognizing the image's character, especially for the image that contains a non-English language or a more complex character, and improving the user interface into a more simple and easy-to-use user interface in order to give the best user experience for the visually impaired people when using the web application.

REFERENCES

[1] Montoya, S. (2018, October 17). Defining Literacy. Retrieved from GAML UIS UNESCO: https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2018/12/4.6.1_07_4.6-defining-literacy.pdf

[2] Novrizaldi. (2021, November 19). Tingkat Literasi Indonesia Memprihatinkan, Kemenko PMK Siapkan Peta Jalan Pembudayaan Literasi Nasional. Retrieved from KEMENKO PMK: https://www.kemenkopmk.go.id/tingkat-literasi-indonesia-memprihatinkan-kemenko-pmk-siapkan-peta-jalan-pembudayaan-literasi

[3] Alyssa Schroer, A. R.-R. (2023, May 19). Artificial Intelligence. What Is Artificial Intelligence (AI)? How Does AI Work? Retrieved from builtin: https://builtin.com/artificial-intelligence

[4] Ramachandran, S. (2023, April). What Is Image Recognition? Retrieved from Nanonets: https://nanonets.com/blog/image-recognition/

[5] Alkhaldi, N. (2022, April 6). How optical character recognition algorithms redefine business processes? Retrieved from itrex: https://itrexgroup.com/blog/how-ocr-algorithms-redefine-business-processes/#header

[6] Rouse, M. (2017, May 9). Speech Synthesis. Retrieved from techopedia: https://www.techopedia.com/definition/3647/speech-synthesis

[7] Inc., E. D. (n.d.). Natural Language Processing. Retrieved from NEURO MARKET: https://neuromarket.ai/natural-language-processing/

[8] J. O. Onaolapo, F. E. (2014, July 2). A Simplified Overview of Text-To-Speech Synthesis. Retrieved from CORE AC UK: https://core.ac.uk/download/pdf/32225276.pdf