FORM 2

THE PATENT ACT, 1970

(39 of 1970)

&

THE PATENTS RULES, 2003

COMPLETE SPECIFICATION

(See section 10 and rule 13)

Title of the Invention:

DYNAMIC VOICE GENERATION SYSTEM AND METHOD

Applicant:

(a)	Name	:	DR. D. Y. PATIL INSTITUTE OF TECHNOLOGY PIMPRI PUNE
(b)	Nationality	:	Indian
(c)	Address	:	Dr. D. Y. Patil Unitech Society's Dr. D. Y. Patil Institute of Technology Main Campus, Sant Tukaram Nagar, Pimpri, Pune, Maharashtra – 411018

Preamble To The Description

Complete

The following specification particularly describes the invention and the manner in which it is to be performed.

Description (Description shall start from next page)

Claims not applicable for provisional specification. Claims should start with the preamble – "I/We claim" on separate page)

Date and Signature (to be given at the last page of specification)

Abstract of the invention (to be given along with complete specification on separate page)

DYNAMIC VOICE GENERATION SYSTEM AND METHOD

FIELD OF INVENTION

The present invention relates to the field of optical character recognition technology, in particular Text-to-Speech (TTS) technology, specifically focusing to convert written text from various sources—such as PDFs and images—into natural-sounding audio.

BACKGROUND OF THE INVENTION

US9305543B2 relates to a text-to-speech (TTS) processing technique that converts text from electronic documents into audio output. While this approach offers several methods for generating audio files that include speech and contextual audio cues, but it has notable limitations. For instance, the method primarily focuses on basic text parsing and conversion without addressing the nuances of user personalization. Additionally, it lacks comprehensive support for extracting text from diverse sources such as images and pdfs, limiting its applicability. The reliance on embedded markup tags can also complicate the integration process for users unfamiliar with such formats.

Moreover, as disclosed in US9305543B2 method above, it does not provide an intuitive mechanism for voice customization, meaning users cannot generate speech that closely resembles their own voice. This lack of personalization can diminish the listening experience, particularly for individuals who benefit from hearing content in a familiar voice. Furthermore, the absence of real-time processing capabilities may hinder the method's usability in dynamic environments where timely audio output is crucial.

Hence, there is a need for a text-to-speech (TTS) system that overcomes the limitations of prior art, such as US9305543B2. This system should be capable of accurately extracting text from diverse sources, including PDFs and images, using advanced OCR technology. Moreover, it enables personalized speech synthesis by allowing users to record and incorporate their own voices. By leveraging deep learning algorithms, the system generates natural-sounding speech, support multiple languages, and offer real-time processing capabilities.

OBJECT OF THE INVENTION

An object of the invention is to provide a text-to-speech (TTS) system and method that enables users to seamlessly convert written text from various sources, including pdfs and images, into high-quality, natural-sounding speech.

SUMMARY OF THE INVENTION

The present invention discloses a text-to-speech (TTS) system and method that transforms written text from various sources, including pdfs and images, into high-quality audio output. This innovative method incorporates optical character recognition (OCR) technology for accurate text extraction and employs deep learning algorithms to produce natural-sounding speech. A key feature is the ability for users to record their own voices, allowing the generated speech to closely mimic their unique vocal characteristics and emotional nuances. The invention also supports multiple languages, making it accessible to a broader audience. By combining these elements, the TTS system enhances accessibility for individuals with disabilities, language barriers, or reading difficulties, providing a versatile and personalized auditory experience.

BRIEF DESCRIPTION OF DRAWINGS

Figure 1 presents the architectural overview of the method.

Figure 2 presents the processing steps of the method.

Figure 3 presents the interaction between a user and the system.

Figure 4 presents the main interface of the method.

Figure 5 presents the menu page of the interface.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides a text-to-speech (TTS) system designed to convert written text from various sources into natural-sounding audio. This innovative solution enhances accessibility and personalization, making it particularly beneficial for individuals with disabilities, language learners, and those with reading difficulties.

The present invention overcomes the disadvantages of the prior art by introducing a sophisticated TTS system that incorporates optical character recognition (OCR) technology for accurate text extraction from a wide range of formats, including pdfs and images. This capability significantly broadens the types of documents that can be converted into speech, enhancing versatility.

Additionally, the present invention allows users to record their own voices, facilitating personalized speech synthesis that mirrors their unique vocal characteristics and emotional expressions. By providing customizable features and an intuitive user interface, the invention addresses the usability issues associated with embedded markup tags, making it accessible for users with varying levels of technical expertise.

Furthermore, the method is optimized for real-time processing, ensuring that text-to-speech conversion occurs swiftly, which is essential for applications in educational and accessibility contexts. The present invention not only enhances the functionality of TTS system but also provides a more engaging, personalized, and accessible auditory experience for a diverse range of users.

In an embodiment of the present invention a Text-to-Speech (TTS) system designed to convert written text from various sources into natural-sounding audio which consists of a text extraction module utilizing optical character recognition (OCR) technology for accurately identifying and extracting text from diverse formats, including PDFs and images, a text-to-speech engine which converts the textual content into audible speech, a voice recording method that enables users to record their own voice for personalized speech synthesis, a voice synthesis module integrating the recorded voice with the extracted text to generate a customized voice model and a user interface facilitating document uploads, text input and voice management.

The present invention discloses text extraction module, which employs optical character recognition (OCR) technology. This module is capable of accurately identifying and extracting text from diverse formats, including PDFs and images. It adapts to a wide range of font styles, sizes, and orientations, ensuring precise text extraction even from visually complex documents. Moreover, the text extraction module supports multiple languages for enhancing the system's versatility for a global audience.

The present invention further discloses, once the text is extracted, it is passed to the text-to-speech (TTS) Engine, which converts the textual content into audible speech. This engine utilizes advanced deep learning algorithms, such as Tacotron 2 and WaveNet, to produce high-fidelity, human-like vocalizations. The TTS engine is designed to generate fluid and expressive speech that closely resembles natural human conversation. Notably, the text-to-speech engine also allows for emotional modulation, enabling users to adjust the emotional tone of the speech output for a more engaging listening experience.

The present invention discloses a voice recording method that allows users to capture their own voice for personalized speech synthesis. This feature facilitates user customization, as individuals can record their voice in a controlled environment, ensuring that the synthesized speech reflects their unique vocal characteristics. The Voice Recording stores recorded voices in a secure database, allowing for easy access and use in future applications.

The present invention discloses a voice synthesis module, which integrates the recorded voice with the extracted text to generate a customized voice model and utilizes this recorded voice to produce custom-made speech. It gives them control over the auditory output that integrates the recorded voice with the extracted text to facilitate the generation of personalized speech. This module creates a voice model based on the user's recordings, resulting in speech that closely resembles the user's vocal characteristics. Additionally, users can adjust the emotional tone of the output, enhancing expressiveness and engagement.

The present invention discloses a user interface which facilitates document uploads, text input, and voice management. Users can upload documents, including pdfs and images, or directly input text for conversion. An integrated voice recording feature allows users to easily capture their voice and manage their voice profiles. The user interface enables customization options for adjusting language settings, emotional tone, and other parameters, tailoring the speech output to their preferences.

In another embodiment of the present invention discloses a method of Text-to-Speech (TTS) system for converting written text into personalized audible speech where the text is extracted from a document using an optical character recognition (OCR) method, processing the extracted text through a deep learning-based text-to-speech engine to generate audible speech, recording a user's voice to create a personalized voice model and then synthesizing the generated speech using the personalized voice model.

Text extraction using OCR (optical character recognition) technology in which the initial step involves extracting textual information from various document types such as pdfs, images, and handwritten notes. This extraction is accomplished using optical character recognition (OCR) technology, which examines the input materials and detects text content. OCR's versatility of OCR allows it to handle diverse font types, dimensions, and alignments, ensuring precise results even when dealing with visually intricate documents.

Figure 1 depicts the architecture of the method works by converting text from PDFs or images into speech. Firstly, the text extraction method uses the Optical Character Recognition (OCR) technique in case of input in image format to pull out the text. Then, this text is sent to our model to turn it into speech. A Voice Recorder is used for users to record their voices, which the model uses to create personalized speech. Therefore, users can hear the text spoken in their voice or by their choice. The method also supports multiple languages and lets users change the voice's tone easily.

1. Start: The process begins with the user interacting with the method through the user interface.

2. User Interface: The user can choose between two options:

   a. Upload PDF/Image: If the user selects this option, the method extracts the text from the uploaded document using Optical Character Recognition (OCR) technology. The OCR module is capable of handling various font styles, sizes, and orientations, ensuring accurate text extraction.
   b. Record Voice: If the user chooses this option, the method captures the user's voice through a voice recording method. This recorded voice is used to create a personalized voice model, which will be used to synthesize speech that closely resembles the user's vocal characteristics.

3. Text Extraction (if required): If the user uploaded a PDF or image, the method proceeds to extract the text using the OCR module.

4. Text-to-Speech (TTS) Engine: The extracted or input text is then processed by the TTS engine, which utilizes deep learning algorithms like Tacotron 2 and WaveNet to convert the text into natural-sounding speech. The TTS engine generates speech patterns, including pitch, duration, and frequency, ensuring that the output sounds natural and fluent.

5. Voice Synthesis (if voice recording was done): If the user recorded their voice, the method integrates the recorded voice with the extracted text to create a customized voice model. This model is used to synthesize speech that closely resembles the user's vocal characteristics.

6. Play Audio: The synthesized speech is then played back to the user through an audio player.

7. End: The process concludes with the user listening to the generated audio.

Figure 2 depicts the detailed processing steps of the method. The input is passed through different stages: voice recording, text extraction, and text-to-speech conversion. The speech synthesis

stage uses advanced algorithms such as Tacotron 2 or WaveNet to generate natural-sounding audio. The final output can be played through an audio player or stored in database for later use.

1. Start: The process begins with the user uploading a file (PDF or image) to the method.

2. Extract Text from File:

a. PDF: If the uploaded file is a PDF, the method directly extracts the text from the PDF file.
b. Image: If the uploaded file is an image, the method uses Optical Character Recognition (OCR) to extract the text from the image.

3. Process Extracted Text: The extracted text is processed to prepare it for the text-to-speech conversion.

4. Custom Voice Model: The method uses a custom voice model to convert the processed text into speech. This custom voice model allows for personalized speech output.

5. Play or Save Generated Audio: The generated audio can be either played back to the user or saved for future use.

6. End: The process concludes with the playback or saving of the generated audio.

This process flow outlines the steps involved in converting text from uploaded files into speech, with a focus on utilizing a custom voice model for personalized audio output.

Figure 3 discloses the interaction between the user and system. The user initiates multiple actions within the system, beginning with the option to upload a PDF or image. This input will be processed, and text will be extracted. Another interaction allows the user to record their voice, likely for customizing the playback or training the system to use their voice for speech synthesis. After that the system converts the text to speech, transforming the extracted or uploaded text into an audio output. Finally, the system provides a playback phonic feature, allowing the customer to listen to the generated speech.

1. User: The user interacts with the text-to-speech system.

2. Upload PDF/Image: The user can upload a pdf or image file to the system.

3. Record Voice: The user can record their voice using the system.

4. Convert TTS: The method converts the uploaded text or recorded voice into speech.

5. Playback Audio: The system plays back the generated audio to the user.

This process flow illustrates the interaction between the user and the text-to-speech system. The user can either upload a document or record their voice, and the system processes the input to generate and play back audio.

The system's key elements are optical character recognition (OCR), text-to-speech (TTS) engine, voice recorder and voice synthesis module—are linked together in the following manner:

1. OCR (Optical character recognition): The OCR component acts as the initial stage by introducing extracted text into the method.
2. TTS Engine: This extracted content is then directed to the TTS engine, which utilizes deep learning models to convert text into vocal output.
3. Voice Recorder: The Voice Recorder captures specific voices from users, which are subsequently employed to customize the speech synthesis. The recorded voice was then combined with the text through the Voice Synthesis Module, ensuring that the final output closely resembled the user's voice.
4. Method Interface: Finally, the method interface enables users to upload documents, record voices, and adjust settings (such as modifying speech tone or pace), providing a smooth user experience.

The system is optimized for real-time processing, ensuring that text-to-speech conversion occurs swiftly and smoothly. This capability is crucial for applications in educational settings, accessibility tools, and other environments where timely audio output is essential.

In another embodiment of the present invention discloses an application for a text-to-speech system consisting a main interface functioning as a user portal and a menu page enabling navigation through a structured catalogue of options, having a Menu Tab (401) for accessing the application's main menu, a "Start Recording" button (402) for initiating voice recordings, a Camera/Gallery button (403) for capturing or selecting images, an "Image" icon (404) for uploading image files, a "PDF" icon (405) for uploading PDF files, a Return to Home Tab (501) for returning to the main screen, a "PDF to Audio" option (502) for converting PDFs into audio, a Settings option (503) for customizing application preferences, a Sync to Google Drive option (504) for backing up files, a Backup File/Restore option (505) for managing file backups, a Trash option (506) for recovering deleted files, a Tutorial option (507) for learning the application's features, a Contact option (508) for customer support, an About PDF to Audio option (509) for providing information on the conversion feature, a Recordings option (510) for managing recorded audio and a Previous Files option (511) for accessing previously used files.

Figure 4 discloses the main interface of the method, it functions as a user portal, delivering immediate access to core features through a minimalist design. Upon launch, users encounter options for initiating recording or accessing camera and gallery functions. These selections facilitate swift capture of new media or retrieval of existing files, streamlining the user experience.

1. Menu Tab (401): The user taps on the menu tab to access the main menu of the application.

2. Start Recording (402): The user taps on the "Start Recording" button to initiate a new voice recording.

3. Camera/Gallery (403): The user can tap on the camera or gallery icon to capture a new image or choose an image from their gallery.

4. Upload Image (404): The user can tap on the "Image" icon to upload an image file.

5. Upload PDF (405): The user can tap on the "PDF" icon to upload a PDF file.

This figure outlines the user interaction with the application's main screen, allowing the user to initiate a new recording or upload media files for processing.

Figure 5 discloses the menu page of the interface showcasing a structured catalogue of its core functions, allowing users to navigate through various offerings with ease. As shown in figure 5, the interface was designed to provide convenient access to different operations, featuring separate areas dedicated to each capability.

1. Return to Home Tab (501): The user taps on the "Return to Home Tab" option to go back to the main screen of the application.
2. PDF to Audio (502): The user can tap on the "PDF to Audio" option to convert a PDF file into an audio file.
3. Settings (503): The user can access the settings menu to customize various aspects of the application, such as language preferences, audio quality, and other options.
4. Sync to Google Drive (504): The user can enable the option to automatically back up files to google drive.
5. Backup File/Restore (505): The user can manually back up files or restore previously backed-up files.
6. Trash (506): The user can access the trash bin to view and recover deleted files.
7. Tutorial (507): The user can access the tutorial to learn how to use the application.

8. Contact (508): The user can contact customer support or provide feedback.
9. About PDF to Audio (509): The user can access information about the "PDF to Audio" feature.
10. Recordings (510): The user can access and manage previously recorded audio files.
11. Previous Files (511): The user can access and work with previously used files.

This process flow outlines the various options and features available to the user within the application's menu.

In conclusion, this innovative TTS system significantly advances current technology by combining OCR for effective text extraction, personalized voice synthesis, and advanced neural networks for natural-sounding speech. It is designed to enhance accessibility, support multi-language needs, and provide a customizable auditory experience, making it an invaluable tool for a diverse range of users.

Advantages of the Present Invention:

- Personalized Speech Output: The system allows users to record their own voices, resulting in speech that closely resembles their unique vocal characteristics. This personalization enhances user connection and comfort.
- Natural Sound Quality: Utilizing advanced deep learning algorithms, the TTS engine produces fluid and human-like vocalizations, making the listening experience more pleasant and engaging.
- Emotional Expressiveness: Users can adjust the emotional tone of the speech output, allowing for greater expressiveness and making it suitable for various contexts, such as storytelling or educational materials.
- Multi-language Support: The system can recognize and process text in multiple languages, broadening its applicability and making it useful for a diverse audience.
- Accurate Text Extraction: The integration of optical character recognition (OCR) technology enables accurate extraction of text from various formats, including pdfs and images, improving the system’s versatility.
- Real-Time Processing: The TTS system is optimized for swift operation, ensuring rapid text-to-speech conversion, which is crucial for real-time applications in educational and accessibility settings.
- User-Friendly Interface: The intuitive interface allows users to easily upload documents, input text, and manage voice recordings, making the system accessible for individuals with varying levels of tech-savviness.

- Enhanced Accessibility: By providing an auditory alternative to reading, the system significantly benefits individuals with visual impairments, reading difficulties, or language barriers, promoting inclusivity.
- Customizable Features: Users have the flexibility to modify parameters such as language, tone, and speed, allowing them to tailor the auditory output to meet their specific needs and preferences.

EXAMPLE

1. File Upload or Text Input: The process begins with the input of a pdf, image file, or text manually entered the system. This initial input serves as the foundation for the text-extraction process and forms the basis for all subsequent steps in creating a personalized voice output. The extracted text was then utilized throughout the remaining stages of the customization procedure.
2. Text Extraction: OCR technology is used by the system to extract text from the uploaded file or image. In this manner, the system can now process text and provide an understanding to further convert it into speech; therefore, text and voice can be interfaced in subsequent steps.
3. Voice Recording: In this step, the system captures the voice of the user using a recorder, and thereby gathers all its unique characteristics. The recorded voice is transmitted to the central processing method, after which it is stored in the voice database. Hence, it is used to generate personalized speech output during synthesis.
4. Text-to-Speech (TTS) Engine: The output text is then processed using a text-to-speech engine. Tacotron 2 would process the extracted text and generate a spectrogram. This spectrogram would contain information about the pitch, duration, and other characteristics of the speech. The given text content undergoes reprocessing to prepare it for voice synthesis using the user's recorded speech. Our model was trained on a specific voice, enabling the conversion of written material into spoken words that sound like the user's own voice.
5. Integration of Voices: The method combines the recorded voice with the processed text, which generates a customized speech output that reflects the user's voice. This necessitates a process of integration to ensure that the generated speech aligns with both textual content and inherent characteristics of the recorded voice sample. WaveNet would take the spectrogram produced by Tacotron 2 and generate the final audio waveform. This

step would produce highly natural-sounding speech that closely matches the user's voice characteristics.

6. Playback and Customization: The method can monitor the generated audio and adjust various aspects, such as intonation and frequency. Customization features offer enhanced choices, individualization, and adaptability in generating speech that aligns with the user's preferences and desired outcomes.

Without being limited to theory, Google Speak has several limitations that are addressed by this invention. Unlike Google Speak, which only offers a set of predefined voices, this invention allows users to create speech output in their own voice, offering a more personalized and unique listening experience. Additionally, Google Speak lacks the ability to accurately extract text from diverse sources like handwritten notes and images, a gap filled by this invention's advanced Optical Character Recognition (OCR) technology. Google Speak also does not support regional languages such as Hindi and Marathi, whereas this invention is specifically designed to cater to users who rely on Indic languages. Furthermore, the invention overcomes the complexity of Google Speak’s interface by offering a simple, user-friendly design that requires minimal technical knowledge, making it accessible to a broader range of users.

While the invention is amenable to various modifications and alternative forms, some embodiments have been illustrated by way of example in the drawings and are described in detail above. The intention, however, is not to limit the invention by those examples and the invention is intended to cover all modifications, equivalents, and alternatives to the embodiments described in this specification.

The embodiments in the specification are described in a progressive manner and focus of description in each embodiment is the difference from other embodiments. For same or similar parts of each embodiment, reference may be made to each other.

It will be appreciated by those skilled in the art that the above description was in respect of preferred embodiments and that various alterations and modifications are possible within the broad scope of the appended claims without departing from the spirit of the invention with the necessary modifications.

We Claim:

1. A Text-to-Speech (TTS) system designed to convert written text from various sources into natural-sounding audio comprises: a Text Extraction Module utilizing Optical Character Recognition (OCR) technology for accurately identifying and extracting text from diverse formats, including PDFs and images; a Text-to-Speech Engine which converts the textual content into audible speech; a Voice Recording method that enables users to record their own voice for personalized speech synthesis; a Voice Synthesis Module integrating the recorded voice with the extracted text to generate a customized voice model; and a User Interface facilitating document uploads, text input, and voice management.

2. The Text-to-Speech (TTS) as claimed in claim 1, wherein the Text Extraction Module supports multiple languages for system's versatility.

3. The Text-to-Speech (TTS) as claimed in claim 1, wherein the Text-to-Speech Engine allows for emotional modulation, enabling users to adjust the emotional tone of the speech output.

4. The Text-to-Speech (TTS) as claimed in claim 1, wherein the Voice Recording method stores recorded voices in a secure database.

5. The Text-to-Speech (TTS) method as claimed in claim 1, wherein the User Interface discloses customization options for adjusting language settings, emotional tone, and other parameters, tailoring the speech output.

6. The Text-to-Speech (TTS) method as claimed in claim 1, wherein the Voice Synthesis Module creates a voice model based on user recordings resulting in speech that closely resembles the user's vocal characteristics.

7. A method of Text-to-Speech (TTS) system for converting written text into personalized audible speech, comprising extracting text from a document using an optical character recognition (OCR) method; processing the extracted text through a deep learning-based text-to-speech engine to generate audible speech; recording a user's voice to create a personalized voice model; and synthesizing the generated speech using the personalized voice model.

8. An application for a text-to-speech system comprising a main interface functioning as a user portal and a menu page enabling navigation through a structured catalogue of options, comprising: a Menu Tab (401) for accessing the application's main menu; a "Start Recording" button (402) for initiating voice recordings; a Camera/Gallery button (403) for capturing or selecting images; an "Image" icon (404) for uploading image files; a "PDF" icon (405) for uploading PDF files; a Return to Home Tab (501) for returning to the main screen; a "PDF to Audio" option (502) for converting PDFs into audio; a Settings option (503) for customizing application preferences; a Sync to Google Drive option (504) for backing up files; a Backup File/Restore option (505) for managing file backups; a Trash option (506) for recovering deleted files; a Tutorial option (507) for learning the application's features; a Contact option (508) for customer support; an About PDF to Audio option (509) for providing information on the conversion feature; a Recordings option (510) for managing recorded audio; and a Previous Files option (511) for accessing previously used files.

Dated this the 11th day of December, 2024

Sushil Kumar [IN/PA/1132]
ACCURES LEGAL
Plot No. 15, Sai Enclave, Sector 23,
Dwarka, New Delhi – 110 077, India

DYNAMIC VOICE GENERATION SYSTEM AND METHOD

ABSTRACT OF THE INVENTION

The present invention provides a Text-to-Speech (TTS) system and method that transforms written text from various sources, including PDFs and images, into high-quality, natural-sounding audio. This system leverages Optical Character Recognition (OCR) technology for accurate text extraction and employs deep learning algorithms, such as Tacotron 2 and WaveNet, to deliver human-like vocalizations. A key feature of the invention is the ability for users to record their own voices, enabling the generation of personalized speech that mimics their unique vocal characteristics and emotional tones. Designed with an intuitive user interface, the TTS system supports multiple languages and offers real-time processing capabilities. This innovative solution significantly enhances accessibility for individuals with disabilities, language learners, and those with reading difficulties, providing a versatile and engaging auditory experience.

Dated this the 11th day of December, 2024

Sushil Kumar [IN/PA/1132]
ACCURES LEGAL
Plot No. 15, Sai Enclave, Sector 23,
Dwarka, New Delhi – 110 077, India

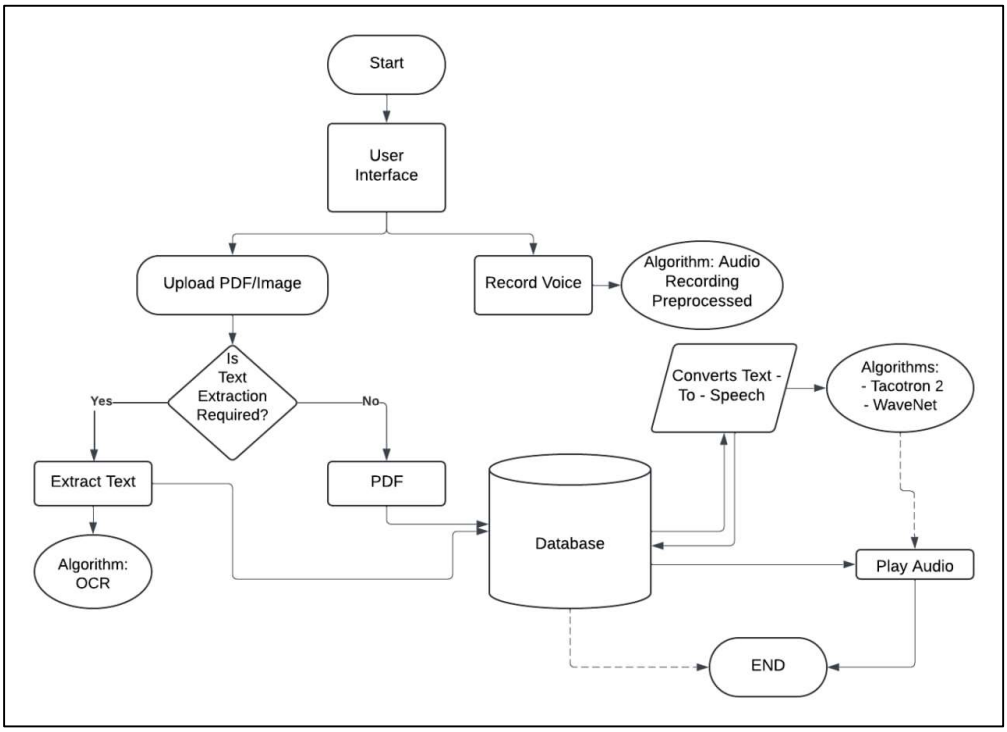

Figure 1

Dated this the 11th day of December, 2024

Sushil Kumar [IN/PA/1132]
ACCURES LEGAL
Plot No. 15, Sai Enclave, Sector 23,
Dwarka, New Delhi – 110 077, India

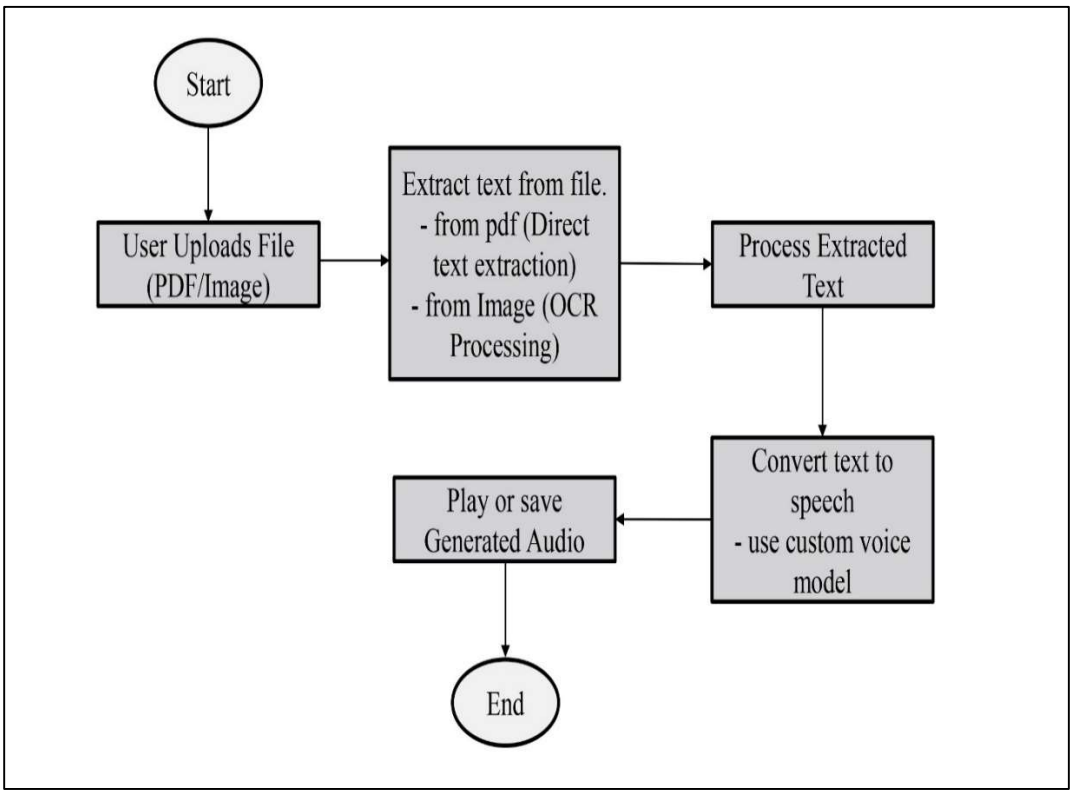

Figure 2

Dated this the 11th day of December, 2024

Sushil Kumar [IN/PA/1132]
ACCURES LEGAL
Plot No. 15, Sai Enclave, Sector 23,
Dwarka, New Delhi – 110 077, India

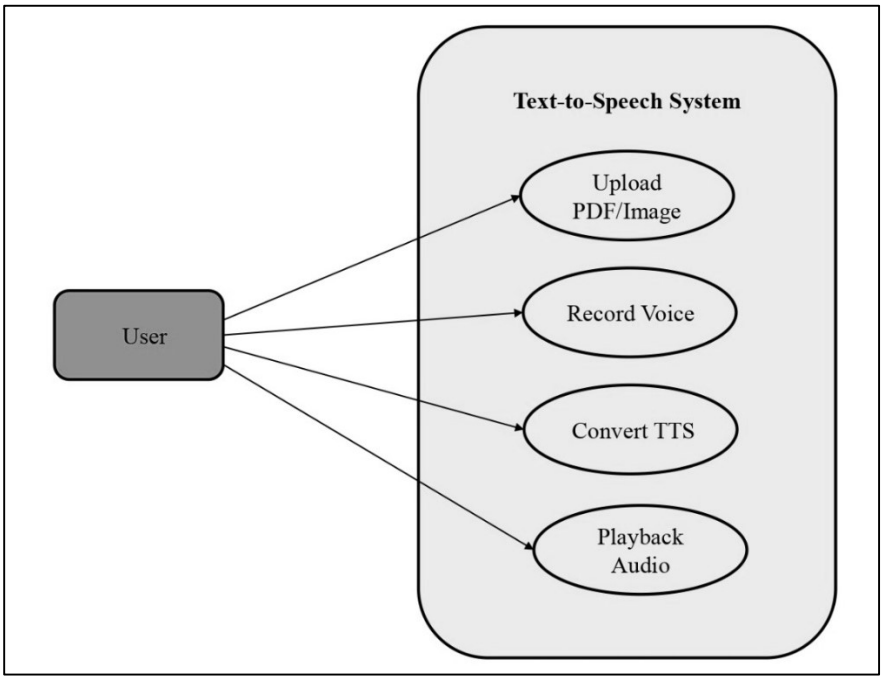

Figure 3

Dated this the 11th day of December, 2024

Sushil Kumar [IN/PA/1132]
ACCURES LEGAL
Plot No. 15, Sai Enclave, Sector 23,
Dwarka, New Delhi – 110 077, India

Figure 4

Dated this the 11th day of December, 2024

Sushil Kumar [IN/PA/1132]
ACCURES LEGAL
Plot No. 15, Sai Enclave, Sector 23,
Dwarka, New Delhi – 110 077, India

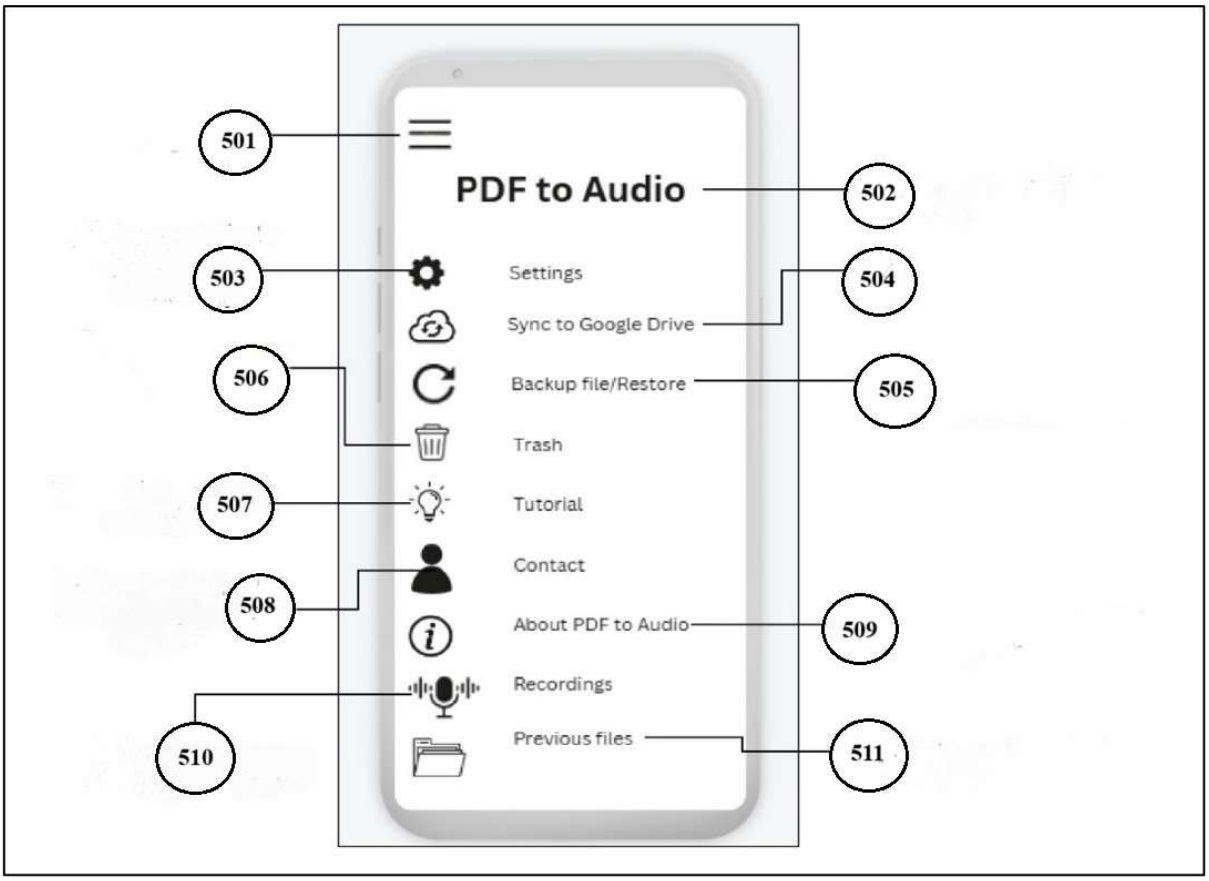

Figure 5

Dated this the 11th day of December, 2024

Sushil Kumar [IN/PA/1132]
ACCURES LEGAL
Plot No. 15, Sai Enclave, Sector 23,
Dwarka, New Delhi – 110 077, India