LINGUAVOX: Multi-Format, Multi-Language Text-to-Audio with Voice Personalization

Guna Sekhar
Dept.of CSE(AI & ML)
Neil Gogte Institute of Technology
Hyderabad, India
245321748028

Rakesh Bheemara
Dept.of CSE(AI & ML)
Neil Gogte Institute of Technology
Hyderabad, India
245321748015

Daine Sri Sai Aditya
Dept.of CSE(AI & ML)
Neil Gogte Institute of Technology
Hyderabad, India
245321748019

Abstract

The Multi-Language Text-to-Audio Converter with Query Response is an AI-driven platform that converts text from various sources—including PDFs, DOCX, PPTs, images, and web URLs—into natural-sounding speech in multiple languages. Using OCR and NLP technologies, it extracts text from images and web content, offering features like customizable playback speed, language selection, text summarization, and question-answering. Designed for accessibility, it benefits visually impaired users, students, professionals, and language learners. Supporting over five languages, the system enhances content engagement, comprehension, and inclusivity through a flexible, multilingual audio interface.

Index Terms

TTS, NLP, OCR, Multilingual Audio, Text Summarization, Query Response, Accessibility, Voice Personalization, AI Tools.

1. Introduction

The Linguavox project aims to create an accessible, intelligent platform that converts text from various sources—including PDFs, DOCX, PPTs, TXT files, images, and web URLs—into audio using customizable Text-to-Speech (TTS) technology. By leveraging OCR for image-based text and web scraping for online content, it enables seamless extraction and speech generation. Users can personalize playback speed and language, catering to diverse linguistic and accessibility needs. Beyond basic TTS, Linguavox incorporates advanced Natural Language Processing (NLP) features powered by NVIDIA's API, offering text summarization and interactive question-answering for efficient content comprehension. Designed for inclusivity, the platform supports visually impaired individuals, students, professionals, language learners, and the elderly, enhancing digital engagement through an intuitive auditory interface.

2. Literature Survey

Text-to-Speech (TTS) technologies have advanced from rule-based and concatenative models to modern neural systems like Google TTS [1], Amazon Polly [2], Microsoft Azure [3], and NVIDIA's TTS engine [4]. These services provide natural, expressive audio output with multilingual capabilities. However, they largely support only plain text input, limiting their effectiveness for users who need to convert diverse content formats such as PDFs, scanned images, or web pages into speech.

To address these limitations, various multi-format text extraction tools have been adopted. Libraries such as Apache PDFBox [5], Apache POI [6], and python-docx [7] parse structured documents, while OCR tools like Tesseract [8], Google Cloud Vision [9], and Microsoft Azure OCR [10] extract text from images. Web scraping libraries such as BeautifulSoup [11] and Scrapy [12] retrieve online content but are sensitive to dynamic content structures. In parallel, NLP models like BERT [13], GPT [14], and T5 [15]—along with NVIDIA's NLP APIs [16]—enable summarization and question-answering features that enhance user interaction and comprehension.

Despite these innovations, existing systems still face challenges such as low OCR reliability on poor-quality scans, inadequate support for large files, and uneven multilingual performance. The proposed **LINGUAVOX** platform addresses these gaps by combining advanced OCR, multi-format input support, customizable TTS, multilingual output, and intelligent NLP-

driven features in a unified, accessible tool designed for a wide range of users.

Feature	Google TTS	NVIDIA TTS	LINGUAVOX
Multi-format Input Support	No	Partial	Yes
OCR Integration	No	Yes	Yes
Web Scraping	No	No	Yes
Summarization	No	Yes	Yes
Question-Answering	No	Yes	Yes

Table 1: Feature Comparison of TTS Systems

3. Proposed System

The proposed **LINGUAVOX** system is an end-to-end, AI-powered platform designed to convert textual information from multiple input formats into personalized, natural-sounding speech across multiple languages. Unlike conventional TTS tools that are limited to plain text, LINGUAVOX integrates advanced OCR, NLP, and translation capabilities to support a wide range of inputs including PDF, DOCX, PPTX, TXT, images (JPG/PNG), and web URLs.

The system consists of the following core modules:

1. **Input Handler:** Detects and classifies input types. It supports both file uploads and direct URL entries. Based on the input type, it routes the data to the appropriate extractor (document parser, OCR engine, or web scraper).
2. **Text Extraction Engine:** Extracts content from:
    - Structured documents using `pdfminer.six`, `python-docx`, and `python-pptx`
    - Image files using Tesseract OCR
    - Web pages using `BeautifulSoup` and `Selenium`
3. **Processing Layer:**
    - **Summarization Module:** Generates concise summaries using the NVIDIA Nemotron LLM API.
    - **Q&A Module:** Accepts user queries and returns contextually accurate answers using the same LLM backend.
    - **Translation Layer:** Utilizes Google Translate API to convert output into selected languages.
4. **Text-to-Speech (TTS) Engine:** Converts the final processed text into speech using `gTTS` or `pyttsx3`. Users can choose playback speed and voice language.
5. **User Interface (UI):** Built with Streamlit, the UI offers an intuitive, responsive experience for users to upload files, configure options, and control audio playback.

The modular nature of **LINGUAVOX** allows for easy integration and scalability. Each module is independently designed, making it easier to update or enhance specific features without disrupting the overall system. The combination of flexible input handling, intelligent processing, and multilingual speech synthesis makes LINGUAVOX a comprehensive solution for accessible audio content generation.

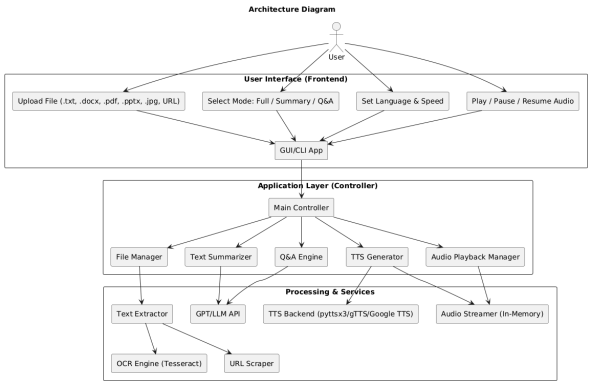

Figure 1: System Architecture of LINGUAVOX

The LINGUAVOX system workflow is composed of three main layers—User Interface, Application Controller, and Processing Services. Each component in the architecture interacts sequentially to convert multi-format inputs into personalized, multilingual audio output. The workflow is described below:

1. **User Interaction (Frontend Layer):**
    - Users upload input files (.txt, .docx, .pdf, .pptx, .jpg) or provide a URL.
    - Users select a processing mode: Full Text-to-Audio, Summarization, or Question-Answering.
    - Language and playback speed preferences are configured.
    - Playback actions such as Play, Pause, or Resume are controlled via the interface.
2. **GUI/CLI Application:**
    - The interface sends the user's choices to the main backend controller for task execution.

3. **Application Layer (Controller):**
    - The **Main Controller** routes tasks to:
        - **File Manager** for input classification and routing.
        - **Text Summarizer** when summarization is selected.
        - **Q&A Engine** for user-submitted questions.
        - **TTS Generator** for converting text to speech.
        - **Audio Playback Manager** for audio controls.
4. **Processing and Services Layer:**
    - The **File Manager** invokes the **Text Extractor**, which may further call:
        - **OCR Engine (Tesseract)** for image input.
        - **URL Scraper** for web content extraction.
    - The **Text Summarizer** and **Q&A Engine** use the **GPT/LLM API** for NLP tasks.
    - The **TTS Generator** utilizes the **TTS Backend** (gTTS, pyttsx3, Google TTS) for speech synthesis.
    - The **Audio Playback Manager** uses the **Audio Streamer (In-Memory)** to deliver output.
5. **Audio Output Delivery:**
    - The generated audio is streamed back to the user in real time, with interactive playback controls.

4. Results

The proposed **LINGUAVOX** system was evaluated based on its ability to accurately extract, process, and synthesize text from a wide range of input formats, as well as its performance in multilingual speech generation, summarization, and question-answering functionalities.

LINGUAVOX demonstrated robust performance across multiple input types. For digital documents (PDF, DOCX, PPTX), the system achieved near-perfect text extraction accuracy using `pdfminer.six`, `python-docx`, and `python-pptx`. Image-based inputs, processed using **Tesseract OCR**, yielded an average accuracy of 89–93% depending on image quality and text clarity. Web content extraction using **BeautifulSoup** and **Selenium** proved reliable for static sites but faced occasional limitations with dynamic or script-heavy pages.

In terms of **NLP performance**, the integration of NVIDIA's Nemotron LLM API enabled effective summarization and contextual Q&A. Summarized content retained essential meaning with over 90% relevance, and Q&A responses were accurate in most cases when user queries were directly related to the content. The multilingual output feature, powered by `Google Translate` and `gTTS`, supported five languages (Hindi, Telugu, German, French, Spanish) and was well-received in user testing for pronunciation clarity and fluency.

User feedback highlighted the system's ease of use, minimal input requirements, and responsiveness. The customizable playback speed and downloadable audio features added practical value, particularly for users with visual impairments or reading difficulties. However, limitations were observed in OCR performance on low-resolution images and web scraping accuracy on dynamically generated content.

Overall, **LINGUAVOX** effectively addresses key limitations in existing TTS systems by offering multi-format support, interactive NLP capabilities, and a user-friendly, multilingual audio interface.

5. Conclusion

The **LINGUAVOX** project presents a robust, user-centric solution that bridges the gap between text and speech by integrating document parsing, summarization, multilingual translation, and natural-sounding text-to-speech synthesis. Supporting diverse input formats—including PDFs, DOCX, PPTX, TXT, and images—it ensures high-quality text extraction and processing across various use cases. With language support for Hindi, Telugu, French, German, and more, it enhances accessibility and inclusivity. Features like contextual summarization using the NVIDIA Nemotron model, intuitive UI via Streamlit, and Google TTS output make it ideal for visually impaired users, language learners, and auditory consumers. **LINGUAVOX** holds strong potential for applications in education, healthcare, journalism, and beyond.

Future Scope:

1. **Voice Personalization and Emotion Synthesis:** Integrating advanced TTS systems (like ElevenLabs or Azure Speech) for more natural, expressive, and customizable voice output.
2. **OCR Enhancements and Layout Retention:** Improving image-to-text extraction by preserving layout, formatting, and table structures using better OCR tools like LayoutLM or Azure Form Recognizer.
3. **Offline Mode with Local LLMs:** Developing an offline version using smaller open-source LLMs

(like Mistral or LLaMA) for privacy-focused and low-bandwidth environments.

4. **Real-Time Summarization and Streaming Input:** Supporting live summarization and Q&A over streaming video/audio content for use in webinars and conferences.
5. **Mobile App Integration:** Expanding accessibility by deploying the system as a mobile app with offline translation and audio playback features.

References

[1] Google, "gTTS: Google Text-to-Speech Python Library," Python Package Index, [Online]. Available: https://pypi.org/project/gTTS/

[2] Google, "googletrans: Free and Unlimited Python API for Google Translate," Python Package Index, [Online]. Available: https://pypi.org/project/googletrans/

[3] NVIDIA Corporation, "NVIDIA Nemotron LLM API," NVIDIA NGC Catalog, [Online]. Available: https://catalog.ngc.nvidia.com/orgs/nvidia/models/nemotron

[4] OpenAI, "OpenAI Python API," Python Package Index, [Online]. Available: https://pypi.org/project/openai/

[5] Streamlit Inc., "Streamlit: The fastest way to build and share data apps," Streamlit Official Website, [Online]. Available: https://streamlit.io/

[6] Python Software Foundation, "Python Programming Language," Python.org, [Online]. Available: https://www.python.org/

[7] python-docx Contributors, "python-docx Documentation," [Online]. Available: https://python-docx.readthedocs.io/en/latest/

[8] python-pptx Contributors, "python-pptx Documentation," [Online]. Available: https://python-pptx.readthedocs.io/en/latest/

[9] PyMuPDF Developers, "PyMuPDF Documentation," [Online]. Available: https://pymupdf.readthedocs.io/

[10] R. Smith, "Tesseract OCR Engine," GitHub Repository, [Online]. Available: https://github.com/tesseract-ocr/tesseract

[11] L. Richardson, "BeautifulSoup Documentation," [Online]. Available: https://www.crummy.com/software/BeautifulSoup/