VocalizeQA: Smart Document Conversion with Interactive Audio and Real-Time Q&A

Bismah Shaikh [1], Sumedh Joshi [2], Manasi Lavekar[3], Vaishnavi Mahale [4], Mr. A. G. Dhamankar[5]

[1]Student, SCTR's Pune Institute of Computer Technology (IT), Pune, Maharashtra, India,
bismahshaikh4@gmail.com
[2]Student, SCTR's Pune Institute of Computer Technology (IT), Pune, Maharashtra, India,
sumedhjoshi463@gmail.com
[3]Student, SCTR's Pune Institute of Computer Technology (IT), Pune, Maharashtra, India,
manasilavekar1@gmail.com
[4]Student, SCTR's Pune Institute of Computer Technology (IT), Pune, Maharashtra, India,
vaimahale74@gmail.com
[5]Assistant Professor, SCTR's Pune Institute of Computer Technology (IT), Pune, Maharashtra, India,
agdhamankar@pict.edu

Abstract

This research paper presents an innovative system designed to enhance the efficiency and accessibility of text-to-speech (TTS) systems for audiobook generation, coupled with a personalized question-answering (QA) mechanism using a retrieval-augmented generation (RAG) pipeline. The system addresses current limitations in TTS, particularly around inference speed and scalability, by leveraging open-source models optimized through techniques such as model pruning, quantization, and hardware acceleration. These enhancements ensure rapid conversion of multi-format documents (PDF, text, HTML) into download- able audiobooks. Additionally, the system integrates a RAG pipeline that processes the same documents to enable real-time, personalized QA. The RAG pipeline extracts and chunks text, images, and tables, storing the data in a vector database for efficient retrieval. Upon receiving a user query, the relevant chunks are retrieved and used as context for a large language model (LLM) to generate precise answers. The dual functionality of this system not only improves user interaction with documents but also expands the usability of TTS and QA technologies in education, research, and accessibility applications.

Keywords: Downloadable Audiobooks, Multi-Modal RAG (Retrieval-Augmented Generation), Personalized QA, and Text-to-Speech technologies.

1. Introduction

In recent years, converting documents into audio has become an essential method for improving accessibility and boosting user engagement across diverse fields. Text-to-Speech (TTS) systems, which convert written content into speech, play a vital role in this process. Their significance is particularly evident in educational settings, research environments, and for individuals with visual impairments or reading difficulties. Audiobooks, for example, have revolutionized the consumption of long-form content, enabling users to multitask and access information on the go.

Despite these advantages, current TTS systems face several limitations, particularly in terms of inference speed, scalability, and personalization, which hinder their effectiveness for real-time or large-scale audiobook generation. While TTS systems provide accessibility, they often lack personalized features, such as the ability to query a document and retrieve tailored information based on user-specific needs.

This research addresses these gaps by proposing a novel system that combines optimized open-source TTS models with a retrieval-augmented generation (RAG) pipeline to deliver a dual-purpose solution: downloadable audiobooks and a real-time question-answering (QA) system. By tackling the key challenges of TTS inference speed and incorporating a personalized QA system, this study aims to enhance the overall user experience when interacting with multi-format documents (PDF, text, and HTML). The primary focus of this work is to improve accessibility, scalability, and personalization within TTS systems for broader applications in educational and research contexts.

The remainder of this paper is organized as follows: Section 2 presents a comprehensive literature survey, discussing relevant research works and their contributions to the field. Section 3 details the proposed methodology, outlining the key components, techniques, and approaches implemented in this study. Section 4 focuses on results and discussions, analyzing the findings, evaluating system performance, and interpreting key observations. Finally, Section 5 concludes the paper by summarizing the key findings and discussing the future scope for further research and improvements.

2. Literature Survey

Table 1. Overview of the Research Examined

Task	ML Algorithms	Attributes	Data Footprint	Traffic Considered	Degree of Classification
"Voice- Assisted Text Summarizer Using NLP", T. N. Charanya and T. C. Sankar, 2023 [11]	NLP algorithms, Entity Extraction	1. Provides hands- free, feedback- driven summarization 2. Accurate text splitting and extraction	Complex text documents	Accessibility for visually impaired users	Summarized content, Application Level
"Development of an Automated Low-Cost Book Scanner and Translator", S. Nawshin et al., 2019 [2]	OCR, Computer Vision (OpenCV)	1. Low-cost book scanner 2. Hardware page-turning mechanism 3. Converts scanned books to audio	Physical book pages	Offline accessibility	E-book creation, Hardware + Software Level
"Knowledge Graph-Based Question Answering System for Remote School Education", L. S. Nair and S. M. K, 2022 [4]	Pretrained Models (e.g., BERT)	1. Integrates knowledge graphs for context-aware QA 2. Handles multi- turn conversations	Knowledge graphs	Educational and remote learning systems	Knowledge-based QA, Semantic Level
"Improving the Question Answering Quality Using Answer Candidate Filtering", A. Gashkov et al., 2021 [6]	Natural Language Processing Features	1. Filters candidate answers for accuracy 2. Continuous updates to ensure precise QA responses	Textual data and queries	QA in knowledge-intensive tasks	Filtered and validated responses, Syntactic Level

"ConvNext- TTS and ConvNext-VC: Fast End-to-End Sequence- to-Sequence Text-to-Speech and Voice Conversion", T. Okamoto et al., 2024 [8]	ConvNext (Neural Networks)	1. Enhances inference speed 2. High-quality sequence-to- sequence voice generation	Speech datasets	High-quality speech synthesis	Text-to-speech transformation, Signal Processing Level
"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", Patrick Lewis et al., 2020 [3]	RAG Pipeline	1. Uses vector databases for content retrieval 2. Enables precise answers with context-based processing	Document chunks, embedding	Real-time query resolution	Knowledge-intensive QA responses, Document Level
"A Review of Deep Learning-Based Speech Synthesis", Yishuang Ning et al., 2019 [9]	Deep Neural Networks	1. Focuses on acoustic modeling 2. Optimizes conversational AI performance	Large-scale speech data	Conversational and interactive applications	Speech synthesis optimization, AI Research Level
“Text-to- Speech Synthesis: A Systematic Review, Deep Learning- Based Architecture, and Future Directions", Fahima Khanam et al., 2022 [5]	Deep Learning Architectures	1. Taxonomy of TTS architectures 2. Discusses datasets and evaluation metrics Outlines future research paths	Linguistic and Audio Synthesis Repositories	Benchmarking TTS systems	TTS synthesis quality assessment, Research Review Level

Table 2. Model Comparison based on Various Features

Feature	Tacotron 2	VITS	Glow-TTS
Architecture	**Autoregressive**: Generates audio step-by-step, where each step depends on the previous one.	**Non-autoregressive**: Generates audio in parallel, resulting in faster synthesis.	**Non-autoregressive**: Parallel synthesis for speed improvements.

Model Size	**Larger**: Requires significant computational resources due to high parameter count.	**Moderate**: A well-balanced model size allowing for efficient processing.	**Moderate**: Smaller model size than Tacotron 2, focused on speed and efficiency.
Inference Speed	**Slow**: Step-by-step autoregressive generation takes significant time.	**Fast**: Non-autoregressive architecture allows faster audio synthesis.	**Fast**: Optimized for real-time or near real-time use cases.
Efficiency	**Low**: Computation-intensive, mainly due to autoregression.	**High**: Combines speed, quality, and lower computational requirements.	**High**: Faster synthesis without significant resource demands.
Output Quality	**Excellent**: Produces human-like audio with minimal artifacts.	**Excellent**: Achieves state-of-the-art audio quality with better naturalness.	**Very Good**: Close to natural audio but might show subtle differences compared to VITS or Tacotron 2.
Use Case	**High-quality speech**: Suitable for tasks requiring natural and expressive speech synthesis, such as audiobooks or podcasts.	**Real-time, high-quality**: Best for high-fidelity applications where both speed and quality are critical, like conversational AI.	**Real-time speech**: Great for scenarios needing quick responses, like chatbots or assistants.
Vocoder Dependency	**Yes (WaveGlow)**: Requires an external vocoder for waveform generation.	**Yes (Integrated Vocoder)**: Vocoder is integrated directly into the model, simplifying the architecture.	**Yes (WaveGlow)**: Relies on WaveGlow or similar vocoder to synthesize raw waveforms.

3. Proposed Methodology

In this segment, the proposed system architecture designed to convert documents (PDFs, PPTs, webpages) into audiobooks while providing a question-answering (QA) system based on document content using a Retrieval-Augmented Generation (RAG) pipeline, has been described. The architecture, as illustrated in Fig. 1, comprises two major components that work in parallel: a text-to-speech (TTS) module for audiobook generation and a RAG pipeline for enabling personalized document-based QA.

Workflow of system:

1. Start:
   The system is initiated. Users are presented with an interface to upload input files for processing.
2. Input the Document:
   The user uploads a document. This document may be in PDF, text, or other supported formats, potentially including content like plain text, images, and tables. The uploaded document serves as the basis for two parallel pipelines.
3. Text-to-Speech (TTS) for audiobook generation.
   Information retrieval and question-answering (QA) with a Retrieval-Augmented Generation (RAG) framework.
4. Preprocessing and Data Extraction:
   The input undergoes preprocessing tailored to its content type:

   a. Text Extraction with OCR or Direct Parsing:
      If the document is unstructured (e.g., PDF with images or scanned pages), Optical Character Recognition (OCR) extracts text and organizes it into a usable format.
      For structured documents (like plain text), data is parsed directly without OCR.

   b. Storage for RAG:
      Extracted data is processed and indexed in a structured format, such as a vector database or knowledge graph, optimized for retrieval and QA tasks.

5. Audiobook Generation with TTS:
   The extracted text is passed to a Text-to-Speech model:
   a. Sentences are converted to speech with natural intonation and clarity.
   b. Audio snippets corresponding to each page or section are generated incrementally.
   c. These snippets are merged into a cohesive audiobook using audio concatenation or streaming methods.
6. Simultaneous Pipelines:
   a. Downloadable Audiobook:
      The merged audiobook is packaged and presented to the user for download.
   b. Personalized QA System:
      The indexed data is made accessible for real-time question-answering.
      When users input queries, the system retrieves relevant information using the RAG pipeline and generates responses.
7. Completion:
   The user is provided with the final audiobook and QA functionalities. The workflow ends. The system will be initiated when the user uploads a document, which serves as the input for both the TTS and QA pipelines. The document may consist of a variety of content types, including text, images, and tables, which are preprocessed to ensure effective downstream tasks. This bifurcation of the input handling allows for simultaneous processing, enhancing system efficiency and reducing latency.

3.1 Document Processing Workflow of Document Processing:

1. Start:
2. The document processing subroutine is initiated upon receiving the document from the main system.
3. Input the Document:
4. The document is divided into manageable units such as pages.
5. Per-Page Processing:

For each page:

   a. OCR Processing:
      If the page contains images or scanned content, OCR extracts readable text. Predefined layouts may guide text parsing for tables and structured content.
   b. Text-to-Speech Conversion:
      The extracted text is input to a TTS model. The model generates a clear and high-quality audio file for the page's content, maintaining a natural tone.
   c. Merge Audio:
      The newly generated audio file is appended to previously processed pages' audio files. This process is seamless, ensuring no abrupt changes between pages.

6. Repeat for All Pages:
7. Steps 3a to 3c are repeated iteratively for all the document's pages.
8. Provide Final Audiobook:
   Once processing is complete for all pages, the final audiobook is made available as a downloadable file.
9. End:
   The document processing ends, and the output is handed back to the main system for final delivery.

3.2 Text-to-Speech (TTS) Module for Audiobook Generation

The aim is to develop a high-performance Text-to-Speech (TTS) system by evaluating open-source models such as FastSpeech [10] and Parler-TTS [7]. Starting with baseline performance experiments, key metrics like inference speed, accuracy, and audio quality will be measured. Based on these results, the best-performing model for further development will be selected. Once the optimal model is chosen, its performance will be enhanced through optimization techniques such as batch processing, OCR for faster text extraction, GPU acceleration, and experimenting with different vocoders. The chosen model will then be used to convert uploaded documents into audiobooks, beginning with the extraction and preprocessing of raw text from formats like PDFs and PPTs. This includes parsing document structures and applying techniques like punctuation normalization. The preprocessed text is fed into the TTS model, which generates natural, coherent speech in real-time. Finally, the generated speech is compiled into an audio file that users can download for easy auditory consumption. Insights from these experiments and optimizations will guide the deployment of an efficient and scalable TTS solution.

3.3 RAG Pipeline for Personalized Document-Based QA

Simultaneously, the uploaded document will be processed through RAG-based QA system [3], allowing users to engage with the document through natural language queries and receive contextually relevant responses. In addition to extracting text, the system also retrieves images and tables embedded within the document, which are crucial for accurately answering questions related to visual data or tabular information. The system summarizes these images and tables to distill essential information, allowing them to be seamlessly integrated into the QA process. These summaries are stored as structured data points for efficient retrieval. The extracted text, along with the summarized visual content, is divided into smaller, contextually coherent chunks, which are then converted into dense vector representations using an appropriate embedding model. These vectorized chunks are stored in a vector database, allowing the system to quickly locate relevant content when responding to user queries. The RAG pipeline combines retrieval and generation in a hybrid framework. When a query is submitted, the system fetches the most pertinent sections of documents from the vector database, which are enriched with contextual information and passed to a large language model (LLM). The LLM generates a well-formed and informative response based on the retrieved data. This personalized response is then delivered to the user in real-time, enabling an interactive Q&A system specific to the content of the uploaded document.

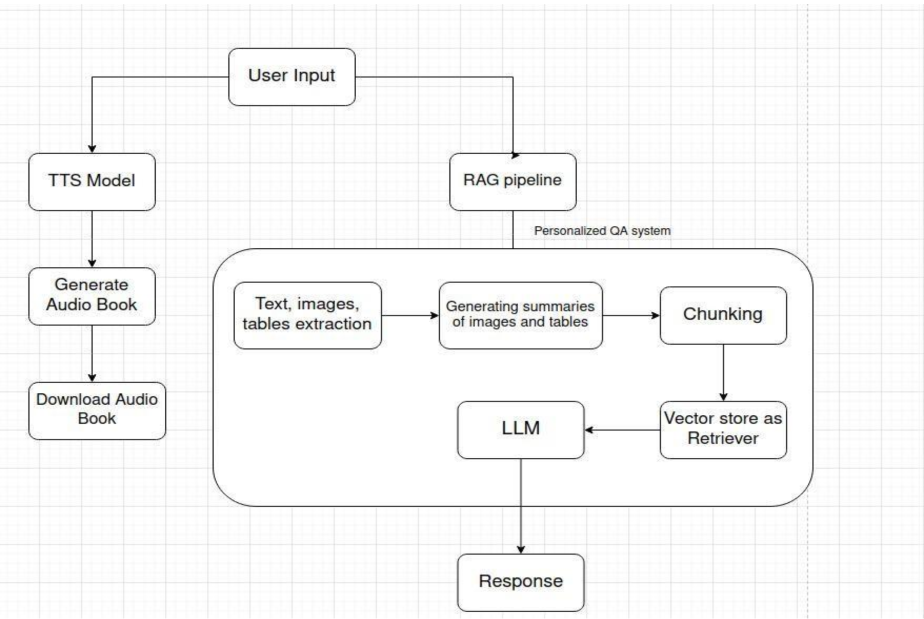

Fig. 1. Proposed System Architecture

3.4 System Output

The final output of the proposed system will consist of two key deliverables. First, the user is provided with a downloadable audiobook version of the document, allowing them to consume the content in an audio format at their convenience. Second, users can engage with the uploaded document through an interactive Q&A system, where they can pose questions and receive personalized, contextually accurate responses based on the document's content. These two deliverables offer a comprehensive solution for both auditory consumption and interactive document exploration.

4. Results and Discussions

4.1 Performance Evaluation Metrics

1. Mean Opinion Score (MOS)

A qualitative assessment method in which human listeners evaluate the synthesized speech's quality using a 1 to 5 scale (5 for Excellent, 4 for Good, 3 for Average, 2 for Poor, and 1 for Bad). The final score is calculated by averaging all individual ratings. For example, if five listeners rate a speech sample as 4, 5, 3, 4, and 4, the MOS would be 4.0, indicating good quality. While MOS is widely accepted, its subjectivity and dependence on human evaluators make it resource-intensive and potentially inconsistent across different evaluation sessions.

$$\textbf{MOS} = (\boldsymbol{\Sigma}\ \textbf{Individual Scores}) / (\textbf{Number of Ratings}) \quad \textbf{(1)}$$

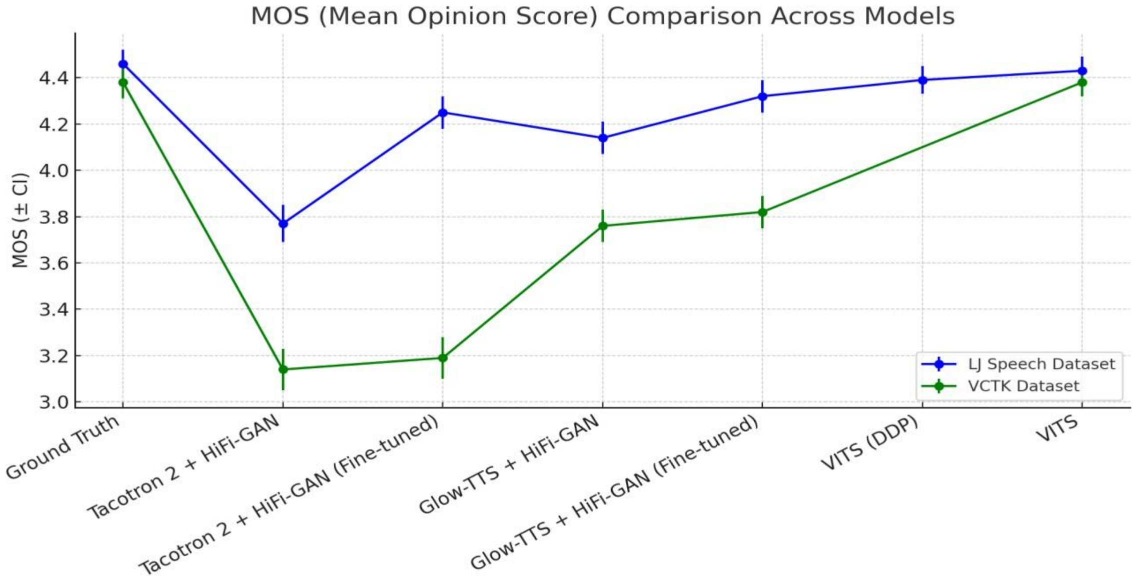

Fig. 2. MOS comparison

2. Word Error Rate (WER)

WER evaluates the precision of speech recognition by determining the least number of word edits (substitutions, deletions, insertions) required to convert the reference text into the hypothesized text. For example, if the reference is "the cat sat on the mat" and the hypothesis is "the cat sat the mat", one deletion ("on") occurs, leading to a WER of 16.67%. A lower WER signifies higher accuracy. This measure is especially valuable for assessing how well synthesized speech is transcribed back into text in terms of both clarity and accuracy.

$$\mathbf{WER = (S + D + I) / N \times 100\%} \quad (2)$$

Where: S = Number of substitutions, D = Number of deletions, I = Number of insertions, N = Total number of words in reference text

3. BLEU Score

BLEU evaluates the standard of machine-generated text by drawing parallels with human references, focusing on n-gram precision. The score includes a brevity penalty to prevent very short outputs from scoring high. BLEU scores range from 0 to 1, with 1 being perfect alignment with references. While originally designed for machine translation, it's adapted for TTS evaluation to measure text preservation accuracy. This metric is valuable for assessing how well the TTS system maintains the original text's structure and meaning.

$$\mathbf{BLEU = BP \times exp\ (\Sigma\ wn \times log\ pn)} \quad (3)$$

Where: BP = Brevity Penalty = min (1, exp (1 - r/c)), r = reference length, c = candidate length, wn = weight for n-gram precision, pn = n-gram precision

4. Real-Time Factor (RTF)

RTF measures processing efficiency by comparing the time taken to generate speech with the duration of the generated audio. An RTF less than 1 indicates faster-than-real-time processing (e.g., RTF = 0.2 means generating 10 seconds of audio takes 2 seconds), while RTF greater than 1 indicates slower-than-real-time processing. This metric is crucial for real-time applications and system optimization. Hardware capabilities, model complexity, and implementation efficiency directly influence RTF.

$$\mathbf{RTF = Processing\ Time / Audio\ Duration} \quad (4)$$

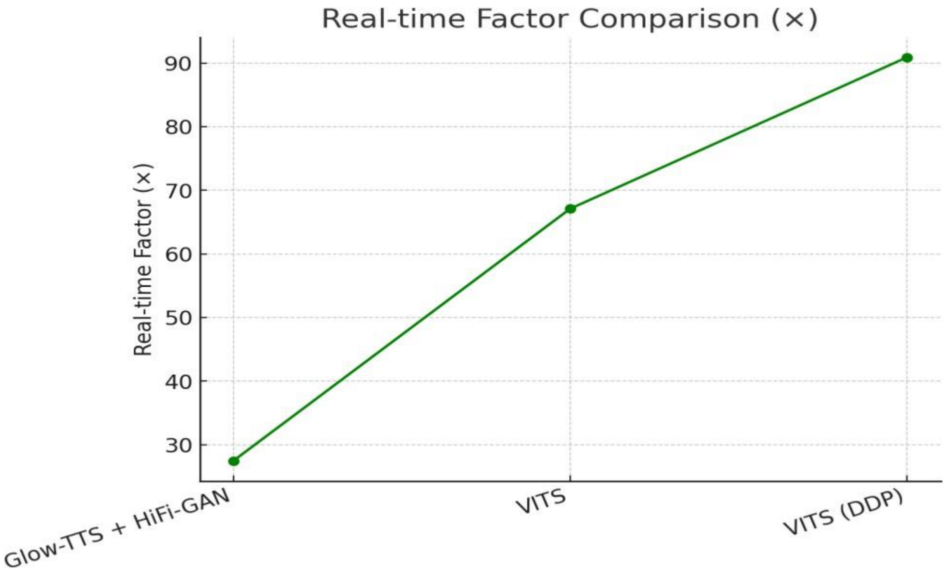

Fig. 3. Real-time Factor comparison

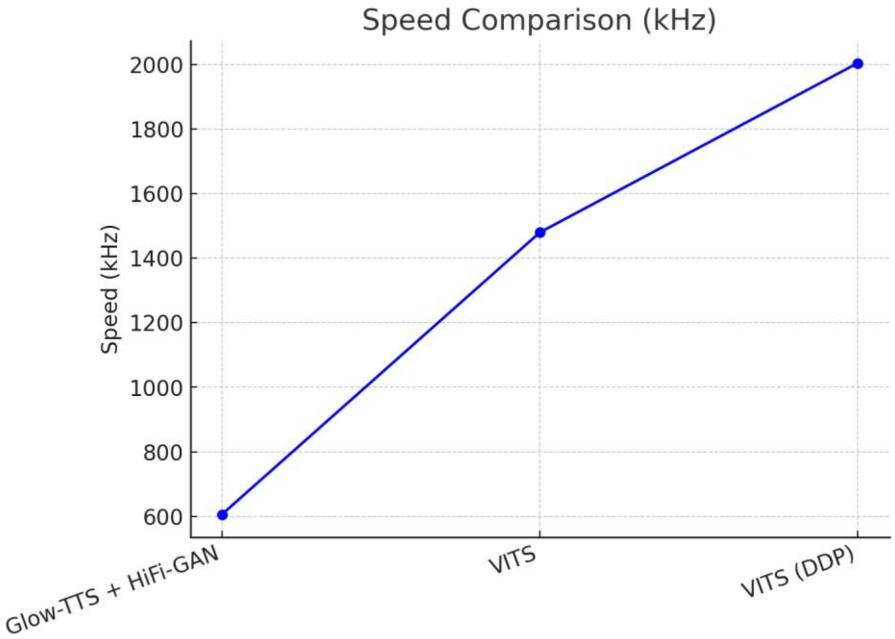

Fig. 4. Speed comparison

5. Conclusion & Future Scope

The future development of VocalizeQA can focus on several key areas for improvement and expansion, including multilingual support to handle multiple languages for broader inclusivity, and complex visual data interpretation to process and convey charts, graphs, and other intricate visuals using advanced computer vision techniques. Enhanced audiobook narration could introduce context-aware features, such as varied vocal tones and structured narration based on document formatting, while an adaptive QA system could learn from user interactions to deliver more accurate responses and manage multi-turn conversations. Scalability and performance optimization will ensure real- time use of text-to-speech (TTS) and vector retrieval as the system grows, while personalization features like adjustable narration styles and enhanced accessibility options will improve the user experience. Finally, platform integration with tools such as Google Drive or Learning Management Systems (LMS) could boost adoption in educational and professional environments, making the system more versatile and powerful in meeting diverse user needs.

This research paper proposes a novel system architecture designed to transform documents (e.g., PDFs, PPTs) into audiobooks while integrating an interactive question-answering (QA) system based on document content. By leveraging a combination of text-to-speech (TTS) technology and a Retrieval-Augmented Generation (RAG) pipeline, the system enables users to access documents audibly and engage with an intelligent, document-specific QA experience. Experimental results demonstrate [16] that Tacotron 2, Glow-TTS, and VITS achieved Mean Opinion Scores (MOS) ranging from **3.13 to 4.45**, with VITS consistently outperforming other models and closely approximating the ground truth. These findings highlight that advanced models like VITS can achieve high-quality, natural-sounding speech synthesis with enhanced accuracy. Additionally, the stochastic duration predictors and end-to-end training employed in VITS allow the generation of high-fidelity audio samples while maintaining low inference times, ensuring practical efficiency.

The proposed methodology presents a comprehensive framework for processing diverse document formats, extracting and synthesizing text, images, and tables, and delivering personalized responses to user queries. Although still in the proposal stage, the approach holds significant potential for improving document accessibility and fostering user interaction with complex content. With further development, this system could transform how users engage with documents, providing both convenience and deeper insights through advanced interactive technologies.

References

[1] K. S. Sri, C. Mounika and K. Yamini, "Audiobooks that converts Text, Image, PDF-Audio & Speech-Text: for physically challenged & improving fluency," *2022 International Conference on Inventive Computation Technologies (ICICT)*, Nepal, 2022, pp. 83-88, doi: 10.1109/ICICT54344.2022.9850872.

[2] S. Nawshin, N. Hossain, S. K. Das, A. Shafin Mohammad Mahdee Jameel, J. D. Mela and S. Islam, "Development of an Automated Low-cost Book Scanner and Translator," *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, 2019, pp. 1-5, doi: 10.1109/ICASERT.2019.8934635.

[3] Lewis, Patrick & Perez, Ethan & Piktus, Aleksandara & Petroni, Fabio & Karpukhin, Vladimir & Goyal, Naman & Küttler, Heinrich & Lewis, Mike & Yih, Wen-tau & Rocktäschel, Tim & Riedel, Sebastian & Kiela, Douwe. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 10.48550/arXiv.2005.11401.

[4] L. S. Nair and S. M. K, "Knowledge Graph based Question Answering System for Remote School Education," *2022 International Conference on Connected Systems & Intelligence (CSI)*, Trivandrum, India, 2022, pp. 1-5, doi: 10.1109/CSI54720.2022.9924128.

[5] Fahima Khanam, Nadia Afrin Ritu, Farha Akhter Munmun, Aloke Kumar Saha, and Muhammad Firoz Mridha, "Text to Speech Synthesis: A Systematic Review, Deep Learning Based Architecture and Future Research Direction," Journal of Advances in Information Technology, Vol. 13, No. 5, pp. 398-412, October 2022.

[6] A. Gashkov, M. Eltsova , A. Perevalov and A. Both, "Improving the Question Answering Quality using Answer Candidate Filtering based on Natural-Language Features," *2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, Chengdu, China, 2021, pp. 635-642, doi: 10.1109/ISKE54062.2021.9755382.

[7] Dan Lyth and Simon King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations", 2024.

[8] T. Okamoto, T. Toda, Y. Ohtani and H. Kawai, "Convnext-TTS And Convnext-VC: Convnext-Based Fast End-To-End Sequence-To-Sequence Text-To-Speech And Voice Conversion," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 12456- 12460, doi: 10.1109/ICASSP48485.2024.10446890.

[9] Ning, Yishuang & He, Sheng & Wu, Zhiyong & Xing, Chunxiao & Zhang, Liang-Jie. (2019). A Review of Deep Learning Based Speech Synthesis. Applied Sciences. 9. 4050. 10.3390/app9194050.

[10] Yi Ren, Yangjun Ruan, Tao Qin, Sheng Zhao, Xu Tan, Zhou Zhao and Tie-Yan Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech", 2019.

[11] T. N. Charanya and T. C. Sankar, "Voice Assisted Text Summarizer Using NLP," *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, Chennai, India, 2023, pp. 1-5, doi: 10.1109/ICDSAAI59313.2023.10452662.

[12] P. Agrawal, K. Dhage, K. Sharma, I. Sharma, N. Rakesh, and G. Kaur, "Speech-to-Text Conversion & Text Summarization," presented at the First International Conference on Technological Innovations & Advance Computing, Bali, Indonesia, 2024, pp. 536-541, doi: 10.1109/TIACOMP64125.2024.00094.

[13] Y. Xiao, L. He, X. Wang and F. K. Soong, "Improving Fastspeech TTS with Efficient Self-Attention and Compact Feed-Forward Network," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 7472-7476, doi: 10.1109/ICASSP43922.2022.9746408.

[14] K. Bapat, S. Paygude and R. Chitti, "Machine learning-based text to speech conversion for native languages," presented at the 2023 14th International Conference on Computing Communication & Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-5, doi: 10.1109/ICCCNT56998.2023.10308393.

[15] A. T. Sujan, K. V. Ujwal Karanth, Y. R. Thanay Kumar, S. Joshi, K. P. Asha Rani and S. Gowrishankar, "Breaking Barriers in Text Analysis: Leveraging Lightweight OCR & Innovative Technologies for Efficient Text Analysis," 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2023, pp. 359-366, doi: 10.1109/ICACRS58579.2023.10404305.

[16] Kim, Jaehyeon, Jungil Kong, and Juhee Son. "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech." *arXiv preprint*, 2021, arxiv.org/abs/2106.06103.

[17] Ito, K. The LJ Speech Dataset. https://keithito. com/LJ-Speech-Dataset/, 2017.

[18] Veaux, C., Yamagishi, J., MacDonald, K., et al. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017.