www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 546-548

e-ISSN : 2583-1062
Impact Factor : 7.001

AUTOMATED PDF TEXT TRANSLATION AND SPEECH GENERATION USING GOOGLE COLAB

Aditi Dubey[1], Ariba Sheikh[2], Siddhi Mohitkar[3], Mahi Yelne[4], Mohammad Tahir[5]

[1,2,3,4]Department of Information Technology G H Raisoni College of Engineering, Nagpur, India.

[5]Center of Excellence in Information Security Department of Information Technology G H Raisoni College of Engineering, Nagpur, India.

DOI: https://www.doi.org/10.58257/IJPREMS44185

ABSTRACT

This research explores the development of an automated system that extracts textual content from PDF files, translates it into a target language, and generates natural-sounding speech output. Using open-source libraries such as PyPDF2, googletrans, and gTTS (Google Text-to-Speech), integrated within the Google Colab environment, the proposed framework provides a scalable and accessible solution for multilingual accessibility of documents. The system enhances information inclusivity by transforming written documents into translated audio, thereby supporting education, research, and accessibility for visually impaired users and non-native language speakers.

Keywords: PDF Processing, Machine Translation, Text-To-Speech, Google Colab, Accessibility.

1. INTRODUCTION

The exponential growth of digital documents necessitates efficient methods for cross-linguistic accessibility and enhanced inclusivity. Traditional PDFs often contain information restricted by language barriers or inaccessible to individuals with visual impairments. While standalone translation and text-to-speech (TTS) tools exist, they often lack seamless integration for end-to-end document accessibility.

Google Colab provides a cloud-based environment for implementing such workflows without local resource constraints. By leveraging machine translation and TTS, this research proposes an automated pipeline that enables:

1) Extraction of text from PDF documents
2) Translation of text into a chosen target language
3) Generation of audio files to deliver the translated content as speech

This integration represents a cost-effective and scalable approach to bridging accessibility gaps.

2. LITERATURE REVIEW

The rapid growth of digital documentation has driven research in automated text extraction, translation, and speech synthesis. Existing tools for document accessibility often operate in isolation, offering either translation or text-to-speech (TTS), but rarely an integrated framework. Recent developments in natural language processing (NLP) and cloud computing have enabled researchers to bridge these gaps.

Several works have explored the use of PDF parsing libraries such as PyPDF2 for extracting text from digital documents, providing a foundation for subsequent processing tasks [1]. Once extracted, the text can be processed through machine translation systems such as Google Translate, accessed via the googletrans Python API, which has demonstrated strong performance in multilingual applications [2]. Despite this, challenges remain in accurately handling idiomatic expressions and domain-specific terminology.

Parallel advancements in TTS technologies have significantly improved the naturalness of synthesized speech. Google Text-to-Speech (gTTS) has been widely adopted due to its ability to generate high-quality audio from translated text [3]. Similar cloud-based services, including Amazon Polly and Microsoft Azure TTS, have been studied for their role in enhancing accessibility for visually impaired users, non-native speakers, and learners. However, limitations in customization and expressive voice generation persist.

Cloud platforms such as Google Colab provide an effective environment for integrating these components into a unified workflow . Prior studies have highlighted the benefits of Colab's resource efficiency, particularly for researchers and educators working without access to high-performance computing infrastructure. Moreover, the literature emphasizes the importance of incorporating Optical Character Recognition (OCR) for scanned PDFs, which remains a limitation in current frameworks.

In summary, prior research demonstrates that while text extraction, translation, and TTS technologies have been studied extensively as independent components, fewer works have proposed integrated systems. The convergence of

these tools within cloud-based platforms represents a promising approach to addressing issues of accessibility, inclusivity, and scalability in digital document processing.

3. METHODOLOGY

A. System Architecture

The system follows a three-phase workflow: sampada.wazalwar@raisoni.net

1) Text Extraction – Using PyPDF2, text is extracted page-by-page from PDF documents
2) Translation – Extracted text is divided into manageable chunks and translated via the googletrans library
3) Speech Generation – The translated text is converted into audio using gTTS

B. Tools and Libraries

- PyPDF2 – Python library for parsing PDF documents
- googletrans – Unofficial Python API wrapper for Google Translate
- gTTS – Google Text-to-Speech library for audio generation
- Google Colab – Cloud-based Jupyter notebook platform

C. Workflow in Google Colab

1) User uploads a PDF file
2) Extracted text is displayed and previewed
3) Language selection via dropdown menu
4) Translated output conversion to MP3 format
5) Audio player embedding and download link generation

4. RESULTS AND IMPLEMENTATION

The system was tested on multilingual documents containing English, Spanish, and Hindi text. Key observations include:

- Accurate text extraction for text-based PDFs
- Translation quality dependent on Google Translate's performance
- Natural speech synthesis through gTTS
- Linear processing time scaling with document length

5. APPLICATIONS

- Education – Native language access to study materials
- Accessibility – Support for visually impaired individuals
- Cross-cultural Research – Facilitation of global knowledge sharing
- Corporate & Government Use – International collaboration support

6. LIMITATIONS AND FUTURE WORK

Current limitations:

- No support for scanned/image PDFs without OCR
- Domain-specific terminology translation challenges
- Limited voice customization in audio generation Future work directions:
- OCR integration (e.g., Tesseract) for scanned PDFs
- Advanced neural machine translation models
- Multi-voice and emotional speech synthesis
- Web/mobile application deployment

7. CONCLUSION

This research demonstrates the feasibility of an integrated system for automated PDF-to-audio translation using Google Colab. The workflow successfully addresses challenges of accessibility and inclusivity while providing a foundation for future enhancements in educational technology and crosslinguistic communication.

8. REFERENCES

[1] PyPDF2 Documentation. https://pypdf2.readthedocs.io/
[2] Googletrans Library. https://py-googletrans.readthedocs.io/

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 546-548

e-ISSN : 2583-1062
Impact Factor : 7.001

[3] Google Text-to-Speech (gTTS). https://pypi.org/project/gTTS/
[4] Google Colab. https://colab.research.google.com/
[5] Tesseract OCR GitHub → https://github.com/tesseract-ocr/tesseract
[6] Mozilla TTS Project → https://github.com/mozilla/TTS
[7] Google AI Research → https://ai.google/research/
[8] A Survey of Deep Learning Techniques for Neural Machine Translation. https://arxiv.org/abs/2002.07526
[9] A Comprehensive Survey of Multilingual Neural Machine Translation. https://arxiv.org/abs/2001.01115
[10] Enhancing the Accessibility of E-Learning Platforms through Text-to-Speech https://dl.acm.org/doi/10.1145/3003733.3003763