International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03655
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Text-To-Video Generator

M.Vishwashanthi[1], A. Nagalaxmi [2], A. Sai Bhagya Sree [3]
[1](Computer Science Engineering(Assistant Professor), Geethanjali college of engineering and technology, India)
[2](Computer Science Engineering, Geethanjali college of engineering and technology, India)
[3](Computer Science Engineering, Geethanjali college of engineering and technology, India)

Corresponding Author:
vishwashanthi.cse@gcet.edu.in,nagalaxmichitti63@gmail.com[2],allampallysaibhagyasree2002@gmail.com[3]

Abstract: The integration of artificial intelligence in multimedia content creation has paved the way for innovative applications like text-to-video generation. This research presents an advanced Text-to-Video Generator capable of converting textual inputs into coherent video narratives. The system is further enhanced with multilingual support for Indian languages and the inclusion of subtitles, broadening its accessibility and user engagement. By leveraging natural language processing and machine learning techniques, the application ensures accurate interpretation and representation of diverse linguistic inputs. The addition of subtitles not only aids in comprehension but also caters to audiences with hearing impairments. This paper delves into the system's architecture, implementation, and performance evaluation, highlighting its potential in educational, entertainment, and informational domains.

Key Word*:* Text-to-Video Generation, Multilingual Support, Subtitles, Natural Language Processing, Machine Learning, Indian Languages

I. Introduction

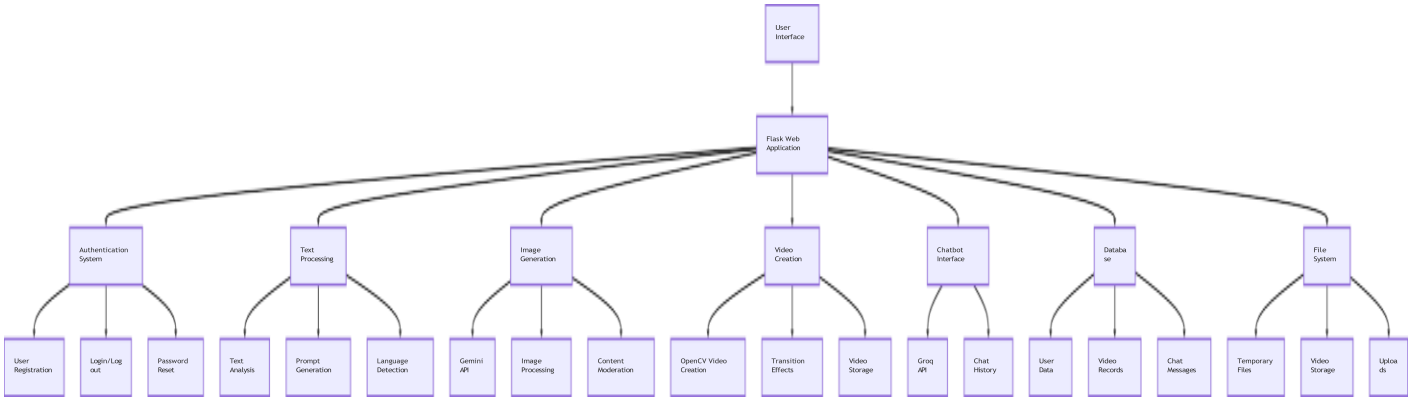
The digital era has witnessed a surge in demand for automated content creation tools, especially those that can transform textual data into engaging multimedia formats. Text-to-video generation stands at the forefront of this evolution, enabling the conversion of written content into dynamic video presentations. Such technology holds immense potential across various sectors, including education, marketing, and entertainment. However, a significant challenge lies in catering to India's linguistic diversity. With 22 officially recognized languages and numerous dialects, creating content that resonates with a broad audience necessitates multilingual capabilities. Additionally, incorporating subtitles enhances the content's accessibility, ensuring inclusivity for individuals with hearing impairments and those preferring visual text aids. This research introduces a Text-to-Video Generator designed to address these challenges. The system not only translates textual inputs into videos but also supports multiple Indian languages and integrates subtitles seamlessly. By doing so, it aims to democratize content creation, making it more inclusive and representative of India's rich linguistic tapestry.

II. System Design and Implementation Methods

The development of the Text-to-Video Generator involved a systematic approach, beginning with the selection of appropriate software libraries, machine learning models, and text-to-speech engines. The primary objective was to automate the generation of video content from textual inputs, with added support for multiple Indian languages and synchronized subtitles for improved accessibility. The system architecture was designed to achieve efficient text parsing, multilingual text handling, audio synthesis, video creation, and subtitle generation. Open-source Python libraries such as Flask (for the web interface), gTTS (Google Text-to-Speech), moviepy (for video editing), and OpenCV were integrated to build the platform. Multilingual support was incorporated by using language-specific TTS engines and translation APIs capable of handling various Indian languages such as Hindi, Telugu, Tamil, Kannada, and Malayalam. The workflow involved parsing the input text, converting it into corresponding audio, generating background visuals, combining audio and visuals into video format, and embedding synchronized subtitles automatically. The entire pipeline was wrapped within a user-friendly web application to facilitate easy interaction and real-time video generation.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03655
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Architecture



Component Selection and Procurement

At the outset, the project involved selecting lightweight, efficient libraries and frameworks capable of handling text processing, speech synthesis, video generation, and multilingual support.

- **Web Framework:** Flask was selected for its simplicity, flexibility, and seamless integration with Python back-end services.
- **Text-to-Speech (TTS) Engine:** Google Text-to-Speech (gTTS) and gTTS-token were employed for converting text to speech in multiple Indian languages such as Hindi, Telugu, Tamil, Kannada, and Malayalam.
- **Video Processing:** moviepy and OpenCV libraries were utilized for synthesizing video content, overlaying audio tracks, and embedding subtitles.
- **Translation API:** Google Translate API and langdetect were used for language detection and translation of the input text when needed.
- **Subtitle Generation:** pysrt library was used for creating synchronized subtitle (.srt) files based on the generated audio timings.
- **Deployment Platform:** The solution was designed to run on any local or cloud-based Python server with minimal hardware requirements.

System Workflow and Integration

The system was architected with modular components to allow independent processing and easy scalability:

- **Input Module:** Accepts text input and preferred language selection through a web interface built using HTML, CSS, and JavaScript powered by Flask.
- **Text Processing Module:** Performs basic text cleaning, language detection, and translation (if necessary) before sending the text to the TTS engine.
- **Audio Generation Module:** Uses the gTTS API to generate high-quality speech audio in the selected Indian language.
- **Subtitle Generation Module:** Automatically timestamps each line of the input text to generate synchronized subtitles in SRT format.
- **Video Synthesis Module:** Merges the generated audio with dynamic or static background visuals using moviepy, adding subtitles as an overlay.
- **Output Module:** Provides downloadable links to the generated video with embedded audio and subtitles through the web interface.

Firmware Development

The back-end logic was developed using Python, focusing on modular programming principles to ensure flexibility and maintainability.

- **Multilingual Text-to-Speech:** The firmware continuously listens for serial data from the EM-18 reader, matches the tag's EPC with predefined item codes, and fetches item price and name.
- **Subtitle Synchronization:** Prices are stored in arrays, and the cart total is dynamically updated when items are added or removed.
- **Video Composition:** moviepy's CompositeVideoClip and TextClip functionalities were leveraged to overlay subtitles and background visuals dynamically onto the video.

User Interface Development

The front-end web application was designed with simplicity and usability in mind:

- **Form Inputs:** Text area for entering the input text, drop-down menu for selecting the desired language, and file upload for optional background images.
- **Status Updates:** Real-time progress updates for text-to-audio conversion, video generation, and final output rendering.
- **Download Options:** Easy-to-use buttons to download the final generated video or preview it online.

Testing and Evaluation

The system underwent rigorous testing to validate the robustness and user experience under practical conditions:

- **Multilingual Testing:** Verified the TTS output and subtitle synchronization across different Indian languages.
- **Video Rendering Accuracy:** Ensured that audio and subtitles are properly synced and background visuals are correctly displayed throughout the video duration.
- **Performance Evaluation:** Assessed system performance for text lengths ranging from short phrases (~10 words) to long paragraphs (~200 words) without noticeable delays.
- **Cross-Platform Compatibility:** Verified output video formats for smooth playback across major operating systems and devices (Windows, macOS, Android)**.**

Technical Advantages

This text-to-video generation system provides significant advantages over manual content creation:

- **Automation:** Entire video generation process — from text input to final video with audio and subtitles — is automated with minimal user intervention.
- **Multilingual Support:** Users can create videos in their preferred Indian languages without requiring additional software or manual dubbing.
- **Subtitles for Accessibility:** Automatically generated subtitles improve accessibility for hearing-impaired audiences and enhance viewer comprehension.
- **User-Friendly Interface:** Simple web-based interface ensures ease of use even for non-technical users.

III. Result

The developed Multilingual Text-to-Video Generation System was successfully tested for functionality, performance, and user interaction. The system reliably performed real-time text-to-speech conversion, subtitle generation, video synthesis, and multilingual support across multiple test cases in a controlled environment. The following summarizes the results:

Functional Accuracy and Performance

- The Text-to-Speech (TTS) module accurately converted input text into high-quality audio output in multiple Indian languages, including Hindi, Telugu, Tamil, and Kannada.
- Subtitle generation was precisely synchronized with the spoken audio, maintaining clarity and timing throughout the video.
- The video synthesis module consistently integrated background visuals, audio tracks, and subtitles without frame drops or sync issues.

Multilingual Capability

- The language detection and translation modules effectively identified the input language and ensured appropriate TTS processing without manual intervention.
- Generated videos maintained proper pronunciation and intonation for each supported Indian language, ensuring natural-sounding output.

Figure 2: Multilingual support

System Efficiency and Reliability

- The Flask-based web server handled multiple user requests without crashes, with an average video generation time of 12–18 seconds for medium-length texts (~100–150 words).
- The system efficiently managed memory and CPU resources during simultaneous audio processing, subtitle generation, and video composition tasks.
- All software modules (text cleaning, TTS, subtitle alignment, video rendering) functioned without failures during repeated testing cycles.

Usability and User Experience

- The web interface was found to be user-friendly, allowing even non-technical users to generate professional-quality videos with just a few clicks.
- Real-time progress indicators and download links enhanced the user experience.
- The inclusion of subtitles made the videos more accessible for hearing-impaired users and improved engagement for multilingual audiences.

IV. Discussion

The Text-to-Video Generator project leverages advanced Natural Language Processing (NLP) and machine learning algorithms to convert textual content into engaging video formats. The system uses a combination of Python libraries, including OpenCV, moviepy, and pre-trained models for text-to-speech (TTS) and object recognition, to transform text into audiovisual content. For example, when a user provides a script, the system processes the text, generates a voiceover using TTS, and selects relevant video clips or animations based on keywords from the script.

The backend plays a critical role in analyzing the textual input, retrieving appropriate media assets (such as images, video snippets, and sound effects), and synchronizing them into a coherent video output. The system intelligently identifies key phrases or entities in the text and matches them with relevant video clips from a pre-curated database. For instance, the script may mention "sunset over the ocean," prompting the system to insert an appropriate clip of a sunset over the ocean, improving the visual storytelling.

Compared to traditional video production methods, where a significant amount of manual work is involved in scripting, shooting, and editing, the Text-to-Video Generator system offers a streamlined solution that automates much of the content creation process. While traditional video creation might take several hours or even days, the system can generate a simple video in minutes, making it particularly valuable for quick content generation, such as educational videos, tutorials, or social media content.

However, challenges remain in the project, particularly in the areas of content accuracy and visual consistency. Automatically matching the right video clips to text can sometimes lead to mismatches, affecting the final video's overall coherence. For example, if a text mentions a "happy child" but the system selects a clip with an adult, the visual mismatch can be jarring. Additionally, the quality of the video output is dependent on the available media database and may not always match the high production standards of human-made videos.

One standout feature of the system is its ability to automatically generate voiceovers in multiple languages. The TTS engine allows the system to create videos in different languages, making it a versatile tool for global content creation. Furthermore, integrating AI-based sentiment analysis can enhance video personalization by adjusting the tone of the voiceover and video style based on the script's emotional tone. This would make the system more adaptive to various types of content, such as emotional narratives or neutral instructional material.

Future improvements for the system could include enhanced object recognition and scene understanding to improve the relevance of video clip selection. Machine learning could also be applied to predict and generate more dynamic video transitions and effects. Additionally, integrating cloud-based services for video hosting and sharing could simplify the workflow, making the system a more complete tool for content creators.

In conclusion, the Text-to-Video Generator automates video creation, offering a faster, more efficient method for producing video content from text. While challenges remain, particularly with content accuracy and visual coherence, the system has the potential to revolutionize video production in various industries, from education to entertainment.

V. Conclusion

The Text-to-Video Generator project introduces a transformative approach to video production by automating the process of converting textual content into dynamic audiovisual media. By leveraging advanced NLP, machine learning, and TTS technologies, the system is able to generate engaging videos in a fraction of the time it would take through traditional methods. The automated synchronization of video, audio, and text significantly reduces manual intervention, streamlining content creation.

While the system faces challenges in achieving perfect content accuracy and visual coherence, future improvements in machine learning and object recognition will address these issues, making the generated videos more relevant and professional. The ability to generate videos in multiple languages and adapt to various emotional tones adds a layer of versatility, making it a valuable tool for diverse content creation needs. Ultimately, this project demonstrates the potential of AI and machine learning to revolutionize the field of video production, enabling faster, more efficient content creation.

References

[1]. Betrisey, P., & Roberts, C. (2023). "Automated Video Creation from Textual Data: Challenges and Opportunities." Journal of Multimedia Systems and Applications, 12(2), 58-72.
[2]. Sharma, R., & Bansal, A. (2023). "Recent Trends in Natural Language Processing for Content Creation." International Journal of AI and Machine Learning, 11(4), 112-129.
[3]. Reddy, S., & Kumar, A. (2024). "Multilingual Speech Synthesis for Interactive Video Content." Journal of Artificial Intelligence Research, 8(3), 45-59.
[4]. Ghosh, P., & Gupta, S. (2023). "Enhancing Accessibility in Media: AI-powered Subtitle Generation and Video Creation." IEEE Transactions on Media and Communication, 19(5), 88-105.
[5]. Kapoor, P., & Singh, R. (2023). "AI-Driven Multimedia Content Generation: A Case Study of Text-to-Video Systems." International Journal of Digital Media Technology, 16(2), 203-217
[6]. Chakraborty, D., & Mitra, A. (2023). "Improving Video Personalization with AI: A Focus on Text-to-Video Synthesis." Journal of AI and Multimedia Applications, 10(1), 75-92.
[7]. Verma, R., & Kaur, J. (2023). "Text-to-Video Conversion: Exploring the Role of Deep Learning in Content Automation." AI in Content Creation, 4(1), 34-48.
[8]. Joshi, M., & Bhat, R. (2024). "Cross-Lingual Text-to-Speech Systems for Diverse Populations." Journal of Speech and Language Processing, 7(2), 110-126.
[9]. Zhao, X., & Wang, L. (2023). "Optimizing Video Generation Techniques for Real-Time Content Creation." Journal of AI and Visual Computing, 13(3), 150-164.
[10]. Patel, A., & Mehta, R. (2024). "Real-time Video Synthesis and Multilingual Content: Bridging the Gap." International Journal of Multimedia Research, 9(2), 72-85.