Multimodal Speech-to-Video Generation Framework for Object Learning in Specially Abled Education

Sandesh PY
Dept. of Computer Science & Engineering
Vidyavardhaka College of Engineering
Mysuru, India
vvce24cse0583@vvce.ac.in

Dr. Hamsaveni M
Dept. of Computer Science & Engineering
Vidyavardhaka College of Engineering
Mysuru, India
hamsaveni.m@vvce.ac.in

***Abstract*— Recent advancements in generative Artificial Intelligence (AI) have created new opportunities for developing interactive and accessible educational technologies. Specially abled learners, particularly individuals with hearing impairments, often rely on visual learning methods to understand object vocabulary and real-world concepts. However, traditional educational tools such as flashcards and static images provide limited dynamic visual representation. This paper presents a multimodal Speech-to-Video Generation Framework aimed at helping specially abled learners by turning spoken or typed object names into short educational videos. The system combines speech recognition, support for multiple languages, prompt generation, and generative AI-based video creation to generate visual demonstrations that match user input. Instead of training custom datasets, the system uses pretrained generative models run in a GPU-enabled environment to create multiple visual scenes for each object. These scenes are merged one after the other to produce a final educational video that can be accessed through a web-based interface.**

The prototype shows that it's possible to change spoken or written input into visual learning materials using generative AI methods. The proposed framework emphasizes the ability of AI-based multimedia creation tools to improve accessibility and engagement in inclusive learning settings.

Keywords— Speech-to-Video Generation, Generative Artificial Intelligence, Diffusion Models, Multimodal Learning, Inclusive Education.

I. INTRODUCTION

Inclusive education aims to provide equal learning opportunities for all learners, including specially abled individuals. Students with hearing impairments often depend on visual learning methods to understand object vocabulary and real-world concepts. However, traditional educational tools such as textbooks, flashcards, and static images provide limited dynamic representation of objects and actions.

Recent advancements in Artificial Intelligence (AI) have enabled the development of systems capable of generating visual content from textual descriptions. Generative models such as text-to-image and text-to-video frameworks can produce realistic visual outputs based on natural language prompts [5]. These technologies provide new opportunities for creating interactive and accessible educational tools.

Text-to-video generation has emerged as an important research area where textual prompts are converted into short video sequences. Diffusion-based generative models have demonstrated promising performance in synthesizing visual scenes aligned with textual descriptions [2]. Such techniques can be utilized to generate the educational videos that visually represent objects.

In this work, we propose a multimodal Speech-to-Video Generation Framework to help specially abled learners. It converts spoken or typed object names into short educational videos. The system combines speech recognition, prompt generation, and pretrained generative models to produce visual scenes that match the input object. These scenes are combined to form a final educational video accessible through a web interface.

II. RELATED TECHNOLOGIES

With recent developments in artificial intelligence and neural networks, machines have increasingly become proficient at creating visual media that is generated from text or verbal input. Generative models, most notably Generative Adversarial Networks (GANs) and diffusion models, have shown greater success in generating images/ videos based on natural language [5].

Multilingual AI frameworks have been developed to convert educational content into multiple languages while generating synchronized visual media. Such systems integrate translation models, speech synthesis techniques, and video generation algorithms to produce multimedia educational materials in regional languages [1]. These approaches help overcome language barriers and improve accessibility for learners who prefer studying in their native language.

Diffusion models have become a widely used technique for generative media synthesis. These models operate by gradually removing noise from random data to generate structured visual outputs. Diffusion-based text-to-video models can generate video frames aligned with textual descriptions while maintaining visual realism [2]. However, maintaining temporal consistency between frames remains a significant challenge in video generation.

AI-driven accessibility systems have also been proposed to improve digital content accessibility for individuals with

hearing impairments. These systems convert spoken language into visual elements such as images, emojis, or graphical cues using deep learning models and speech recognition algorithms [3]. Experimental results show that visual representations significantly improve comprehension and engagement among hearing-impaired users.

Assistive learning platforms integrating speech-to-text, text-to-speech, and sign language translation technologies have been developed to support inclusive education. These systems use natural language processing and computer vision techniques to provide interactive learning experiences for specially abled students [4].

In addition to generative models, cross-modal learning techniques have been used to improve the relationship between textual queries and video content. Cross-modal representation learning maps textual and visual data into a shared semantic space, enabling efficient retrieval and matching of videos corresponding to textual queries [8].

Recent studies have also explored customized video generation techniques that combine textual prompts with structural guidance such as motion information or depth maps. These approaches use latent diffusion models to generate temporally coherent videos while reducing computational complexity [9].

To address the high computational cost associated with video generation, grid diffusion models have been proposed to represent videos as structured image grids. This approach allows text-to-image models to be extended for video generation tasks while reducing memory requirements [10].

Despite these developments, limited research has focused specifically on using generative video technologies for supporting specially abled learners. Therefore, integrating speech recognition, multilingual processing, and diffusion-based video generation into a unified framework remains an important research direction.

III. Proposed System Architecture

The proposed Speech-to-Video learning system follows a modular architecture that converts speech or textual input into short educational videos. The system integrates speech recognition, natural language processing, and AI-based video generation models.

The architecture consists of the following modules:

- User Input Module
- Speech Recognition Module
- Language Translation Module
- Prompt Generation Module
- AI Video Generation Module
- Scene Composition Module
- Web Interface Module

The overall workflow of the system is illustrated below:

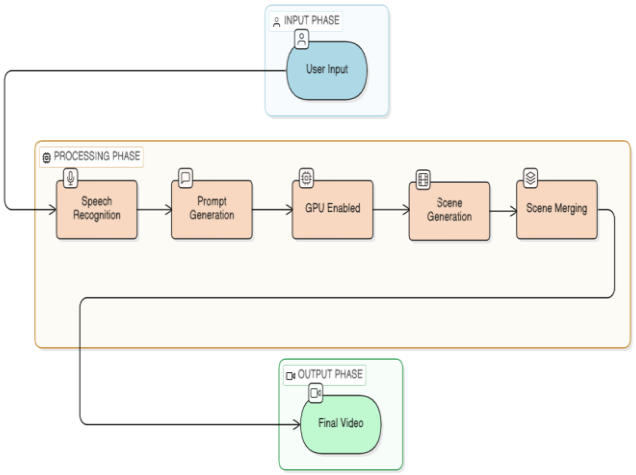

Fig. 1. Proposed Speech-to-Video generation system architecture.

IV. Methodology

A. User Input Module

The system allows users to provide input either through speech or text. The interface captures spoken input using a microphone or typed input through a web interface. This module acts as the interaction layer between the user and the learning system.

B. Speech Recognition Module

If the input is provided in speech format, the system uses Automatic Speech Recognition (ASR) to convert audio signals into textual data. This textual representation is then forwarded to the natural language processing pipeline.

C. Language Translation Module

For the purpose of supporting multilingual users, the system is designed to handle both English and regional language inputs like Kannada. In case the input is given in Kannada script, the system will translate the input into English using neural translation models. This is due to the fact that generative video models are compatible only with English.

D. Prompt Generation Module

The recognized object name is converted into a structured prompt that is used by the AI video generation model. Instead of generating dynamic prompts for each input, the system uses a predefined prompt template to ensure consistency.

Example prompt template:
"A clear educational visual representation of a [INPUT] showing its appearance and movement."

Example:

Input: Dog

Prompt:

"A clear educational visual representation of a dog showing its appearance and movement."

Using a fixed prompt template helps maintain consistency across generated videos and ensures that the generated visuals remain suitable for educational purposes.

E. GPU-Based Video Generation Module

For the generation of the video, a generative AI model is used. This model can either be a diffusion model or a GAN model. Due to the complexity involved in the generative model, the video generation process is carried out in a dedicated GPU environment.

The video generation program is run on a GPU-enabled machine that uses the trained model for inference to generate the visual scene for the given prompt.

Using the GPU for the model speeds up the video generation process. This is due to the high computational power that the GPU can provide for the video generation process.

F. Scene Generation and Composition Module

Instead of creating an entire video in one step, the system creates multiple small scenes that represent different aspects of the object from various points of view. Currently, the system is designed to create different visual scenes for different objects.

Each visual scene represents various aspects such as:

• Changes in the view point

• Changes in the movement of the object

• Changes in the context

These scenes are then combined in sequence to create an entire educational video. The composition of the video is carried out by the video processing pipeline.

This way of creating multiple scenes helps in better visual understanding of the objects and provides a better learning experience for specially abled people.

G. Web Interface Module

The final video output is delivered to the user through a web-based interface. The interface allows users to interact with the system by providing speech or text input and viewing the generated educational video.

The web interface acts as the presentation layer and provides an interactive platform for learners to explore object vocabulary visually.

Due to the computational complexity of generative AI models, the video generation component is executed on a GPU-enabled system while the user interaction and input processing modules operate through a lightweight web interface.

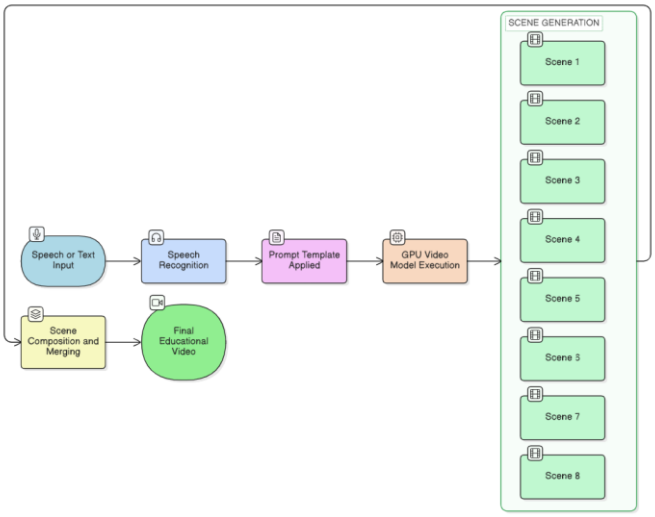

Fig. 2. The Speech-to-Video processing pipeline used for generating educational video.

V. Prototype Implementation

A prototype implementation of the proposed framework for Speech-to-Video generation has been developed. The prototype integrates the various components of the system, including the speech processing component, the natural language processing component, and the AI-based video generation component.

The system has been implemented using a web-based system that allows users to input the names of objects using speech or text input. The input is processed using a pipeline that includes the speech recognition component, the prompt generation component, and the video generation component.

The component of the generative AI that performs the function of creating the video utilizes the GPU environment because of the computational requirements of the deep learning model. The script for the creation of the video runs independently on the GPU server, using the prompt created from the input object to create visual scenes using the generative video model.

In the current implementation, the system creates multiple short visual scenes that represent the input object. The visual scenes are later combined to create a single video for the educational purpose.

The user interface for the system is created using a web application that allows users to easily interact with the model. The user interface allows the user to input speech or text and receive the video output from the system.

VI. Experimental Analysis

The prototype system was tested with multiple object inputs to evaluate the feasibility of generating visual educational content from textual prompts. The experiment focuses on

verifying whether the system can successfully convert object names into visual demonstrations.

Example test cases include objects such as apple, dog, tree, and flower. For each input object, the system generates multiple scenes which are then merged to produce a final video output.

TABLE I. SHOWS EXAMPLE OUTPUTS GENERATED BY THE PROTOTYPE SYSTEM.

Input Object	Generated Scenes	Final Output	Status
Aeroplane	8 scenes	Educational video	Successful
Dog	8 scenes	Educational video	Successful
Lion	8 scenes	Educational video	Successful

The results demonstrate that the system can successfully generate visual learning content corresponding to object names. However, the generation time varies depending on the hardware configuration and GPU availability.

Since diffusion-based video generation models require significant computational resources, the current work focuses on demonstrating the feasibility of the system rather than real-time performance.

Future improvements will focus on optimizing generation time and improving the quality of generated videos.

VII. Results and Discussion

As shown in the experimental evaluation, it is clear that the proposed Speech-to-Video framework is effective in transforming object names into visual educational demonstrations. The generated video provides learners with a better opportunity to understand objects visually.

The use of multiple scenes in generating videos increases the variety of information presented to the learner visually. This increases learner engagement in learning object characteristics, as it is presented in different views.

However, there are a number of challenges in implementing this system. Video generation using generative AI is a task that requires high computational power. The generation of a video takes time depending on the availability of GPUs.

In the current prototype implementation, the system is executed on an edge computing environment with limited hardware resources. As a result, the complete video generation process requires approximately 20 to 30 minutes for a single object input. The system generates multiple visual scenes sequentially, where each scene requires approximately 4 to 8 minutes for generation.

It is expected that the generation time can be significantly reduced when executed in a high-performance GPU-enabled environment. With optimized GPU resources, the total video generation time can be reduced to approximately 2 to 4 minutes, making the system more suitable for near real-time educational applications.

Another limitation associated with the current version of the prototype is the reliance on the pretrained generative models for the creation of videos. As the system is dependent on the pretrained models rather than the training dataset, the diversity and accuracy of the generated videos are dependent on the performance of the generative model.

Overall, the proposed framework proves the possibility of using the available generative AI technologies for converting the input speech or text into visual learning content. The system proves the potential of the AI-based multimedia generation for developing learning tools for specially abled learners.

Fig. 3. Multi-scene generation process showing eight frames extracted from the generated video before scene merging.

VIII. Conclusion

This paper presented a multimodal Speech-to-Video generation framework designed to assist specially abled learners in understanding object vocabulary through visual demonstrations. The proposed system uses speech recognition, multilingual support, prompt generation, and generative AI-based video creation to convert speech or text into an educational video.

The proposed system architecture allows users to input speech or text, which is recognized by the system. This input is then used to create structured prompts. These prompts are then used by pretrained generative models in a GPU-enabled environment to create various visual scenes. These visual scenes are then combined to create an educational video that can be accessed by users using a web-based interface.

The proposed system has been implemented as a prototype that shows the feasibility of converting simple object names into visual learning materials. This is done by creating dynamic visual representations that help learners relate speech or text-based concepts with visual representations, making it easier for specially abled learners.

The proposed system shows the potential of generative artificial intelligence in creating accessible educational materials.

IX. Future Work

Although the current system has successfully proven the feasibility of speech-to-video learning, there are certain aspects that can be improved in the future.

In the future, the system can be improved by focusing on the efficiency of video generation by optimizing the generative model pipeline. Additionally, the system can be improved by incorporating more advanced generative video models that can generate high-quality videos.

Moreover, the system can be improved by incorporating various educational topics that go beyond basic object vocabulary. Additionally, the system can be improved by incorporating more advanced prompt engineering techniques and learning features.

Furthermore, the system can be improved by incorporating various accessibility features that can benefit specially abled people.

X. References

[1] P. Subha, M. Nithya, R. Keerthana, and S. Harini, "AI-Based Multilingual Text-to-Video and Speech Generation System," in *Proceedings of the International Conference on Artificial Intelligence and Smart Systems*, 2025.

[2] Utkarssh Sehgal, Ankit Gupta, and Shashank Sharma, "Text-to-Video Generation Using Latent Diffusion Models: Structural and Temporal Analysis," in *Proceedings of IEEE Madhya Pradesh Conference (MPCON)*, 2025.

[3] Mahdieh Hatami and Mehran Chegini, "Enhancing Digital Content Accessibility for the Hearing Impaired Through AI-Driven Visual Representations," in *Proceedings of the IEEE International Conference on Artificial Intelligence Applications*, 2024.

[4] P. Senthil Kumari, M. Suresh Kumar, R. Karthikeyan, and V. Balamurugan, "Deep Learning-Based Multi-Model Assistive Learning System for Hearing and Visually Impaired Students," in *Proceedings of the IEEE International Conference on Intelligent Systems*, 2023.

[5] Aditi Singh and Rohan Verma, "A Survey of AI Text-to-Image and AI Text-to-Video Generators," in *Proceedings of the IEEE International Conference on Artificial Intelligence and Data Science*, 2023.

[6] Gaganpreet Kaur and Pratibha Sharma, "A Study of NLP and AI-Driven Online Text-to-Video Generation Platforms in Social Media," in *Proceedings of the IEEE International Conference on Smart Computing and Artificial Intelligence*, 2024.

[7] Indira Priya P., R. Harish Kumar, and S. Manikandan, "Transforming Text to Video: Leveraging Advanced Generative AI Techniques," in *Proceedings of the IEEE International Conference on Emerging Technologies in Artificial Intelligence*, 2024.

[8] Jianfeng Dong, Yabing Wang, Xianke Chen, Xiaoye Qu, Xirong Li, Yuan He, and Xun Wang, "Reading-Strategy Inspired Visual Representation Learning for Text-to-Video Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5680–5694, Aug. 2022.

[9] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, and Tien-Tsin Wong, "Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance," *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[10] Taegyeong Lee, Soyeong Kwon, and Taehwan Kim, "Grid Diffusion Models for Text-to-Video Generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.