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

Introduction

- **Goal**: Provide equal learning opportunities for specially abled learners, particularly hyperactive children.
- **Challenge**: Traditional educational tools (flashcards, static images) provide limited dynamic representation of objects and actions.
- **Solution**: A multimodal Speech-to-Video Generation Framework that converts spoken or typed object names into short educational videos.
- **Core Technologies**: Generative AI, Latent Diffusion Models, Natural Language Processing, and Speech Recognition.

Proposed System Architecture

- **User Input**: Accepts speech or textual input.
- **Speech Recognition & Translation**: Converts audio to text.
- **Prompt Generation**: Converts the object name into a structured template.
- **GPU-Based Video Generation**: Uses pretrained generative models.

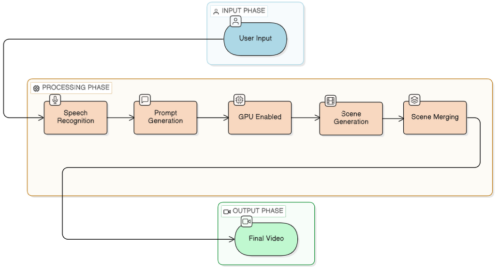

Figure: Proposed Speech-to-Video generation system architecture

Methodology & Core Technologies

- **Speech Processing**: Automatic Speech Recognition (ASR) algorithms convert user audio signals into textual data.
- **Language Translation**: Neural Machine Translation models process regional script inputs (e.g., Kannada) and cross-translate them into English.
- **Prompt Engineering**: Natural Language Processing (NLP) uses predefined, structured templates to maintain consistency in the generated visual characteristics.
- **Video Generation Core**: Pretrained **Latent Diffusion Models** (and Generative Adversarial Networks) synthesize the final video frames mapped to the text prompt.
- **Hardware Acceleration**: The heavy inference pipeline of the generative deep learning models is executed independently on dedicated **GPU-enabled servers**.

Processing Pipeline & Scene Composition

Why Multiple Scenes?

Instead of a single continuous shot, the system merges multiple small scenes.

- Viewpoint changes.
- Object movement.
- Context/background changes.

Pipeline Execution

The scene composition involves a sequential pipeline:

- **Step 1: Contextual Prompting**
  The core object is placed in diverse, educational environments to maintain attention.
- **Step 2: Keyframe Synthesis**
  The Diffusion model generates the base structural grids for each specific view.
- **Step 3: Temporal Smoothing**
  The individual scenes are merged sequentially. This provides an engaging, high-stimulus visual experience tailored for hyperactive learners.

Experimental Analysis & Results

- Tested with objects: *apple*, *dog*, *tree*, *flower*.
- Successfully generated 8-scene educational videos.

Object	Scenes	Status
Aeroplane	8 scenes	Successful
Dog	8 scenes	Successful
Lion	8 scenes	Successful

Figure: Multi-scene generation process showing eight frames

Conclusion

- We presented a novel Speech-to-Video generation framework to assist specially abled learners via visual demonstrations.
- The prototype confirms that using pretrained generative models without requiring custom datasets is a viable approach.
- Dynamic visual representations effectively link speech or text-based concepts to clear visuals, bridging the learning gap.
- This highlights the broader potential of Generative AI in creating accessible educational materials.

Thank You!

Any Questions?