IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

# AI-Driven Virtual Assistant Transforming Voice Into Visuals

**Lavanya B R and Thouseef Ulla Khan**
Department of MCA
Vidya Vikas Institute of Engineering and Technology, Mysuru, India
lavanya200225@gmail.com, thouseef.khan@vidyavikas.edu.in

**Abstract:** *The rapid advancement of artificial intelligence (AI) has revolutionized the way humans interact with machines, enabling more intuitive and human-like communication across multiple modalities. Traditional text-to-image generation systems such as Stable Diffusion and DALL·E have demonstrated remarkable success in converting written prompts into photorealistic or artistic images. However, these systems often rely solely on typed input, which can be inconvenient, especially in scenarios where users prefer natural communication methods such as voice. To overcome this limitation, this project explores the development of a Voice-to-Image Generation System that integrates speech recognition with advanced generative models to produce high-quality images directly from spoken descriptions. The system works by capturing the user's voice input through a speech recognition interface, converting the speech into text using natural language processing techniques, and then feeding the text prompt into a fine-tuned Stable Diffusion pipeline to generate images. The system architecture integrates three main components: voice acquisition and transcription, text-to- image generation, and frontend-backend communication. The frontend enables real-time speech-to-text transcription and user-friendly interactions, while the backend handles prompt management, model inference, and output rendering. This work leverages Stable Diffusion XL (SDXL), a state-of-the-art generative model known for its ability to synthesize detailed and high- resolution images. The dataset used for training and fine- tuning contains paired data of images and text prompts, enabling the model to learn semantic alignment between descriptive phrases and their corresponding visual representations. The project also incorporates strategies such as negative prompting and random seeding to improve image quality and variety.*

**Keywords***:* Artificial Intelligence (AI), Generative AI, Voice-to-Image Generation, Stable Diffusion XL (SDXL), Speech Recognition, Natural Language Processing (NLP), Multimodal Interaction, Deep Learning, Accessibility



DOI: 10.48175/IJARSCT-29957