Audio2Art: Project Design - Part 1

# 1. System Overview

Audio2Art is a two-stage AI system that transforms audio prompts into visual art. The system processes voice input through an Automatic Speech Recognition (ASR) model to generate text, which is then fed into a text-to-image diffusion model to produce an image.

# 2. Architecture Diagram Description

Step 1: User uploads a .wav audio file via Streamlit interface

Step 2: Audio is passed to promptgen(), powered by Wav2Vec2, to transcribe audio to text

Step 3: Transcribed text is sent to text2image(), which uses Stable Diffusion to create an image

Step 4: Output image is displayed and downloadable from the web UI

# 3. Components

- Wav2Vec2: Pre-trained ASR model for transcription

- Stable Diffusion: Text-to-image generator from Hugging Face diffusers

- Streamlit: Web interface for user interaction

- Google Colab: Execution environment with GPU support

# 4. Data Flow

Audio (.wav) → Text (via Wav2Vec2) → Image (via Stable Diffusion)

Intermediate formats: waveform arrays, text strings, image arrays

The Streamlit app acts as the controller for inputs and outputs