# Vision Vanguards: A Computer Vision-Based Karaoke Experience

Team Members: Athulya Ganesh, Stephanie Mullins, Rob Kelly
Advisor: Jillian Aurisano

# Goals

Our team wishes to implement a web-based karaoke application that utilizes a mixture of Computer Vision and audio processing techniques. Our aim is for the user experience to feel like a blend of Just Dance and Karaoke, allowing them to both sing and perform poses to a given song.

- Some examples of use cases that we wish to cover for consumers:
    - Enable users to access the application and utilize both the audio and video portions of the application to perform a wide variety of songs
    - Allow users to save their score via a leaderboard and compare it with their peers
    - Permit users to visit the website, pick a song, and use only the video portion of the application, which would allow them to enjoy the game without worrying about singing
    - Allow for users to access the application and only perform using the audio portion of the application, allowing them to perform without the need for a video camera

# Intellectual Merits

- Integration of Computer Vision and Audio Processing
  - Seamless integration of computer vision and audio processing technologies
  - Enables immersive karaoke experience combining singing and dance routines
- Gesture and Skeletal Tracking
  - Innovative implementation of gesture recognition and full-body skeletal tracking
  - Enhances user interaction and experience with intuitive controls
- Scoring Algorithm
  - Development of a composite scoring system based on audio and visual inputs
  - Evaluates both singing accuracy and dance movements for comprehensive feedback
- Web Application Development
  - Utilization of web technologies for scalability and accessibility
  - Frontend and backend frameworks ensure a robust and user-friendly platform
- User-Centric Design
  - Tailored to diverse user personas including casual singers and shy performers
  - Prioritizes user engagement and satisfaction for an enjoyable karaoke experience

# Broader Impacts

- Entertainment
  - Provides a fun outlet for game night
  - Practice singing and dancing in a nonconventional way
- Collaboration
  - Brings together friends and family with a fun activity
- Technological Innovations
  - Introduces another genre of usage for Computer Vision technologies, focusing more on a consumer-friendly product as opposed to more commercial usages (hospitals, agriculture, etc.)

# Design Specifications

- Project Overview
  - Develop an integrated karaoke and movement web application
  - Combine computer vision and audio processing technologies
  - Objectives: Gesture and skeletal tracking, audio pitch accuracy
- Computer Vision Tasks
  - Research gesture recognition and skeleton tracking
  - Implement hand signal recognition and tracking
  - Develop skeleton detection and tracking algorithms
  - Ensure correctness of skeletal tracking for graded scoring

# Design Specifications (cont.)

- Audio Processing Tasks
  - Research and select audio processing libraries
  - Implement chosen library and integrate it
  - Develop audio to text transcription and pitch detection
  - Assess lyrical and pitch accuracy for composite scoring
- Web Interface Tasks
  - Research frontend/backend technologies and database options
  - Create wireframes and prototypes using Figma
  - Set up development environment and infrastructure
  - Develop web app structure, navigation, and features
  - Implement scoring, leaderboard, and user management

# Design Specifications (cont.)

- Combined Tasks:
    - Integrate audio processing into the web app
    - Combine computer vision with audio and website
    - Implement algorithm for composite performance scoring

# Technologies

- Computer Vision
    - OpenCV
        - Camera input and video manipulation
    - MediaPipe
        - Pose estimation of the camera input
- Audio Processing
    - Librosa
        - Implemented primarily for extracting pitches from audio tracks
    - PyAudio
        - Used for recording and opening audio files
    - PyGame
        - Utilized for loading and playing music tracks
    - SpeechRecognition
        - Employed for transcribing audio tracks to text

# Technologies (cont.)

- Web Interface
    - React JS Framework
        - Used to create the frontend and backend of the web interface
    - Firebase
        - Used for authenticating users as well as maintaining a list of users and high scores

# Milestones

| | October | November | December | January | February | March |
|---|---|---|---|---|---|---|
| **Computer Vision** | Research software and begin environment setup | Complete gesture recognition. Start full body still images skeletal detection | Finish still image full body detection. Begin work on live video with skeletal detection | Implement recognition of human body movements with inputted movements | Finish full body skeletal movements and scoring of user movements | Integrate Computer Vision feature into web application |
| **Audio Processing** | Research audio processing libraries | Implement chosen audio processing library | Research audio-to-text transcription and pitch detection methods | Implement lyrical and pitch accuracy features | Compute composite accuracy algorithm and begin integration with web application | Finish integration with web application |
| **Web Interface** | Determine choice of technology, UI design, and environment setup | Finish web application skeleton and database setup | Create performance interfaces, lyrics module, and gesture guidance | Implement scoring, restart, leaderboard, and user management pages | Testing and refinement of web application | Integrate web application with CV and audio processing |

# Results

- Completed
  - Computer vision logic (pose estimation)
  - Audio processing logic
  - 75% of website properly set up
  - User authentication logic
- Work in progress
  - Hook computer vision and audio portions into web app
  - Complete Firebase database implementation for leaderboard
  - Implement all components for game demo
  - Hardcoding poses for game to correlate to specific songs

# Challenges (Computer Vision)

- Computer Vision
  - OpenPose: Installation problems and compatibility issues with laptop
    - Solution: Tried to debug installation problems and looked at documentation. Ended up moving to another library because of computer hardware compatibility
  - MediaPipe: Finding right methods within library, performance slowness
    - Solution: Researched documentation and tried different pose estimation approaches within MediaPipe's library

# Challenges (Audio Processing)

- Audio Processing
  - Pitch matching algorithm worked, but did not account for different singing tones; notes sung would be correct but register as incorrect
    - Solution: Implementing an octave-matching algorithm allowed for different tones to be taken into account with scoring
  - Copyright and integration issues with Spotify API did not allow for us to properly implement as wide of a variety of songs as we intended
    - Solution: We decided to instead focus on a couple of hard-coded songs that users could perform, and as the project grows in the future, this will be further fleshed out and copyright problems will be worked out

# Challenges (Web Interface)

- Web Interface
  - Issues with firebase connecting to the web application
    - Solution: Debug Firebase Integration thoroughly by checking SDK installation, configuration settings, and authentication setup, using Firebase documentation for guidance
  - Responsive design, especially when dealing with multiple complex components

    Solution: Implement Responsive Component Design using a Mobile-First approach, Flexbox, CSS Grid, Media Queries, and modular component breakdown, while ensuring accessibility considerations are met through testing across devices