Real Time Transcription and Subtitles

Joseph Adelman, Ari Singer, Eli Kersey, Tony Tang

Introduction

- The primary goal of our project is to transform how individuals with hearing impairments engage in conversations and social interactions.
- Original idea - smart glasses that are designed to enable users to "see" spoken words in real-time. When a user looks at someone who is speaking, the text would appear above their head
- Current Product – An application that can take in a camera feed, to then detect the number of people on the screen, as well as detect who is speaking. This application can take in a microphone input, then send that information to the whisper API to receive written text. We also incorporated google translate, which auto detects the spoken language and translates the text into the language of the user.

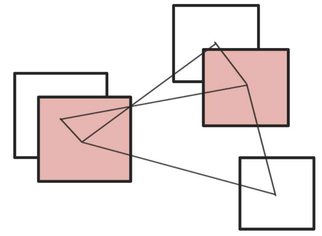
Background

- The motivation for this project was to create a tool that allows people with hearing impairments be able to focus on the speaker, as well as understand what they are saying
- Our original idea was to incorporate this project with smart glasses, such as the Microsoft Hololens. These glasses use a clear display and mixed reality to display digital components on the real world
- For our current application, the best use case would be on a zoom meeting. Closed captioning on zoom puts the text on the bottom of the screen, rather than near the speaker
- We built this product to test how useable this idea is in practice. Real world subtitles seem like a great idea, but it test computer capabilities to translate correct, place text, and detect speakers.
- This project also tests the users ability to read text that is constantly moving around the screen, as most people are not sat in one place when they speak

Design

- **OpenAI Whisper**: provides speech-to-text services for transcription
- **Google Translator:** translate transcription into desired language
- **OpenCV**: Facial recognition and tracking to continuously the ‘head count’ each frame. Track points around the lips of each face, to then determine who the current speaker is based on maximum movement over the previous 10 frames

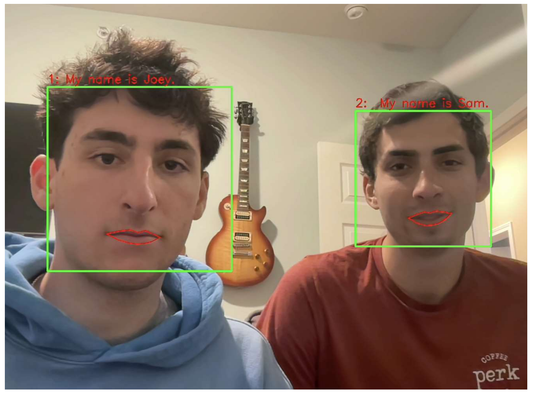
- **White**: Frame 1 Faces
- **Red**: Frame 2 Faces
- Find Shortest distance from old faces to new faces, set new faces as closest pair



Project Prototype

OpenCV Facial Recognition and Mouth Movement Tracking:

- Our algorithm updates with each video frame, connecting the new faces on the screen to an old face that was in the previous frame (nearest neighbor)
- It simultaneously finds the distance between the upper and lower lip, and stores that value in a Face object
- To determine a speaker, the displace between the the top and bottom lips are summed over the last 10 frames
- It manages phrases spoken by each speaker, ensuring the text displayed is associated with the correct speaker.
- For each detected face/speaker combination, it displays the transcribed text near the speakers face in the video feed.



Results

Analysis

- We had users watch conversations of 1-2 people with no sound, and pretend as if they were on a video call to conduct user testing
- 90% of users felt as if they understood the conversation, and could use this in real time
- 60% of users said that this for of subtilties were more enjoyable compared to standard subtitles, with 20% feeling no difference, and 20% feeling as if this form was worse than standard
- The main issues that users stated was that text either impaired their view or moved too much since the text moved with the speaker
- The main features that was missing were too “make the text more readable”, or “add different text colors per speaker”

Conclusion

- Although this product seems very desirable, any issues with the application can cause major issues for all users
- For this product to be useable, no words can be put above someone who is not speaking
- We chose red for the text as it is more readable in most setting rather than white or black, but we need to add a text background, which trades off readability for view obstruction
- We need to update the speaker detection and add user independent phrases rather than sending all input as one phrase to whisper to ensure correct text placement

Future Works

- Find better ways to detect faces and which person is speaking when a video feed has multiple people
- Allow for multiple audio inputs to be able to do transcription with multiple speakers
- Connect this to allow screen grab, or interface with applications such as zoom
- Expand the language capabilities of the translation to include a broader range of languages and dialects.