Multimodal Person Identification through the Fusion of Face and Voice Biometrics

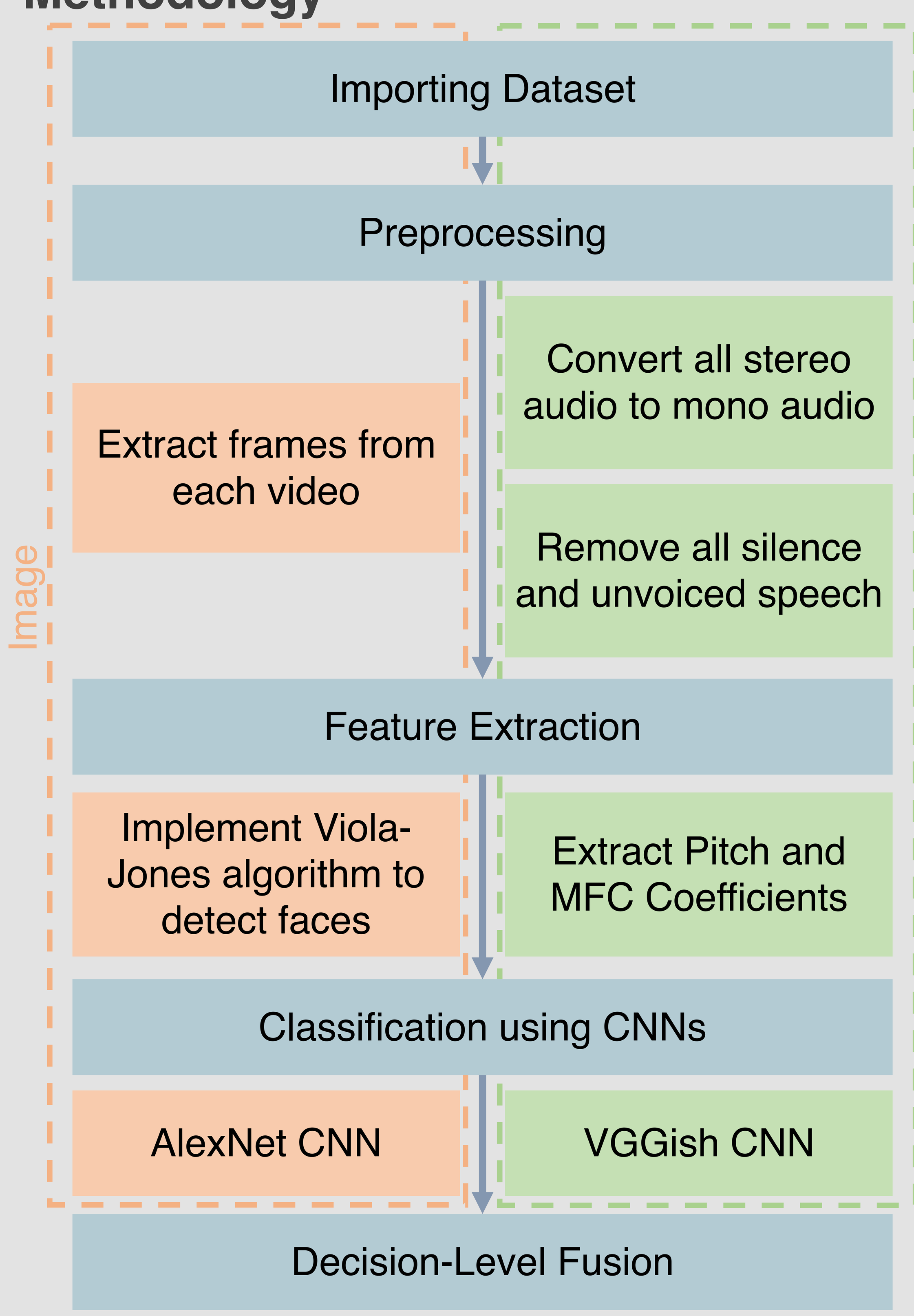
Marwa Saleh, Advisor: Ismail Jouny

Department of Electrical and Computer Engineering, Lafayette College, Easton, PA 18042, USA

Introduction

- Person identification is used every day in a variety of applications from access control and security cameras to social media and face ID on smartphones.
- The process of person identification involves different biometrics such as face, voice, fingerprint, etc.
- However, there are still various limitations to their efficiency which makes them incompletely reliable.
- In public settings, like that of a museum, using one biometric alone becomes challenging due to factors like background noise, overlap of people's faces, varying angles and/or distances from the camera, as well as the recently introduced challenge of face masks.

Methodology



Dataset

Michigan State University Audio-Video Indoor Surveillance (MSU-AVIS) Dataset: [1]

- 50 Subjects (16 females, 34 males)
- Image data variations include:
  - Indoor illumination
  - Facial expressions
  - Pose & distance relative to the camera
- Audio data variations include:
  - Indoor reverberations
  - Background Noise
  - Distance from the microphone

Feature Extraction

Viola-Jones Algorithm:

1. Divide image into squares
2. Calculate delta of sum (shaded) and sum (unshaded)
3. Identify any Haar-like features
4. Crop the $227 \times 227px$ square where a face was identified

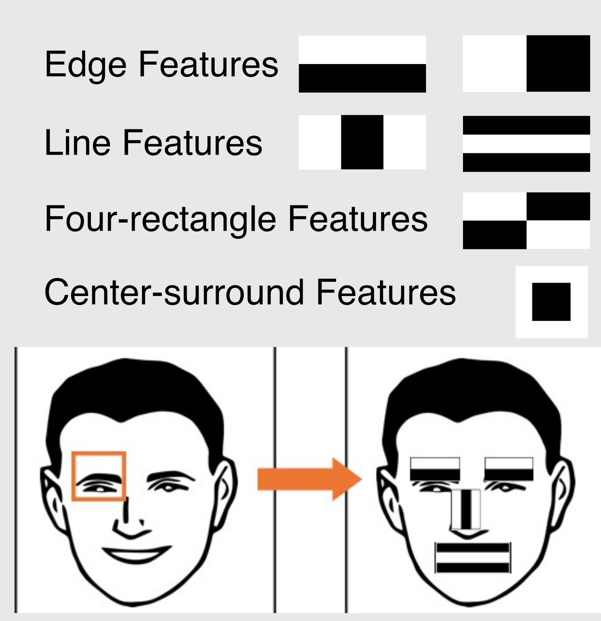


Figure 1. Demonstration of the Viola Jones Algorithm [2]

Pitch:

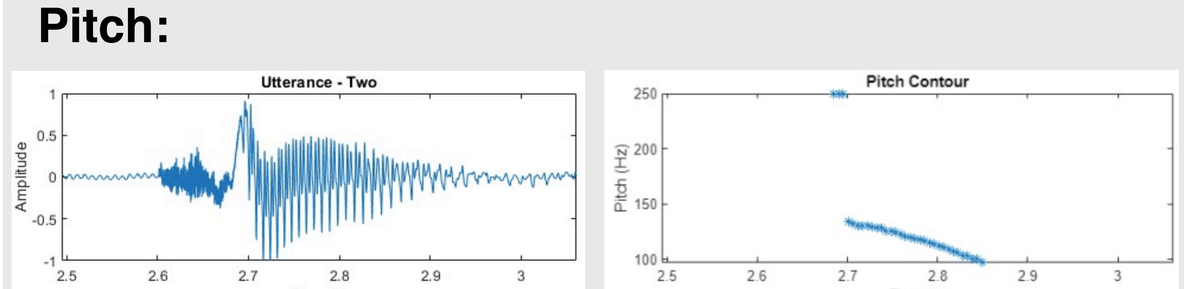


Figure 2. Extracting pitch from the voiced speech portions of an audio [3]

Mel Frequency Cepstral Coefficients (MFCCs):

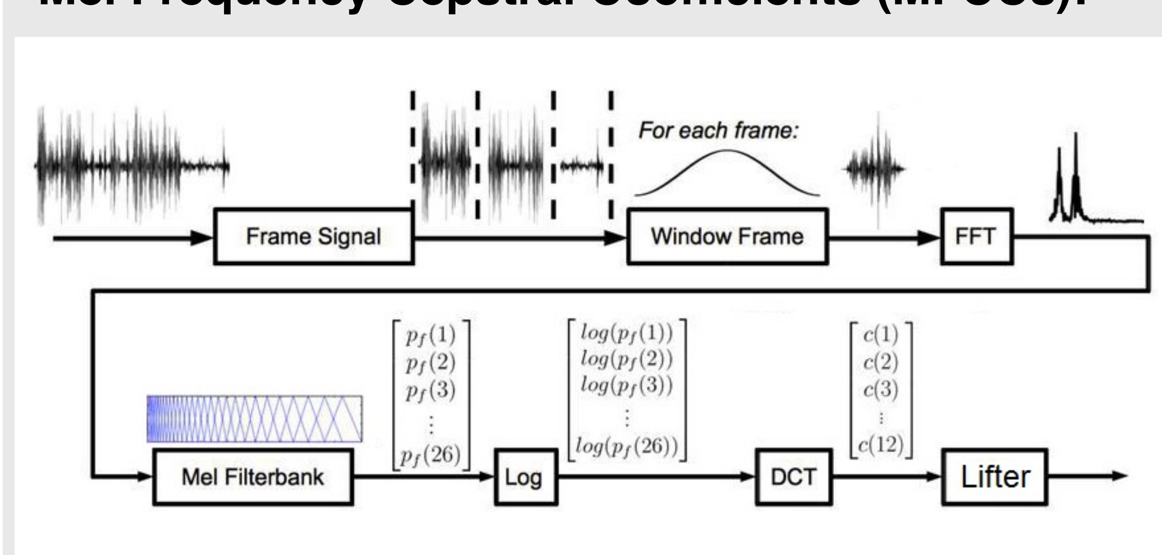


Figure 3. Block diagram of the Mel Frequency Cepstrum [4]

Convolutional Neural Networks (CNNs)

Image CNN:

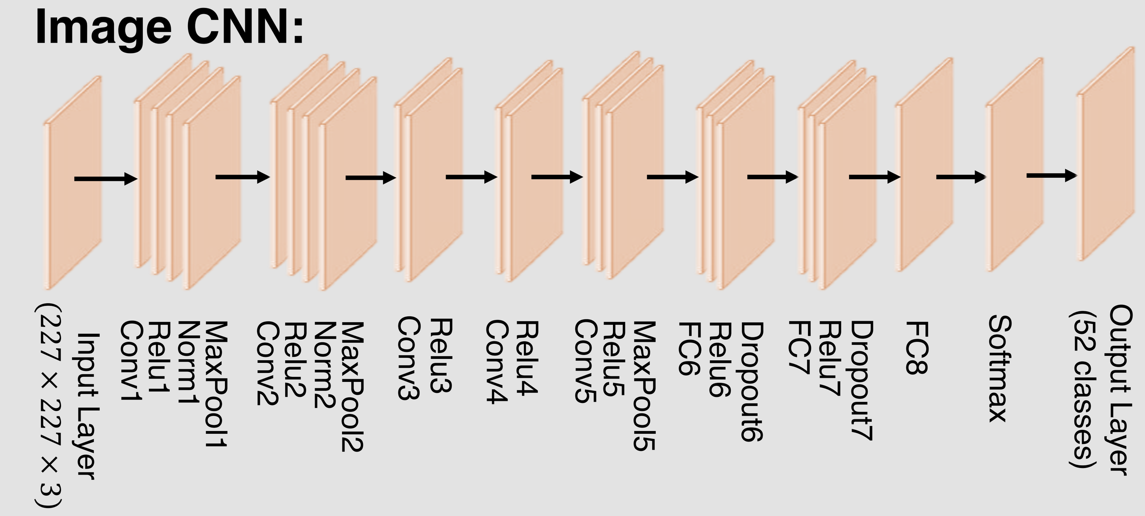


Figure 4. Layers of AlexNet CNN after adjusting them to my problem

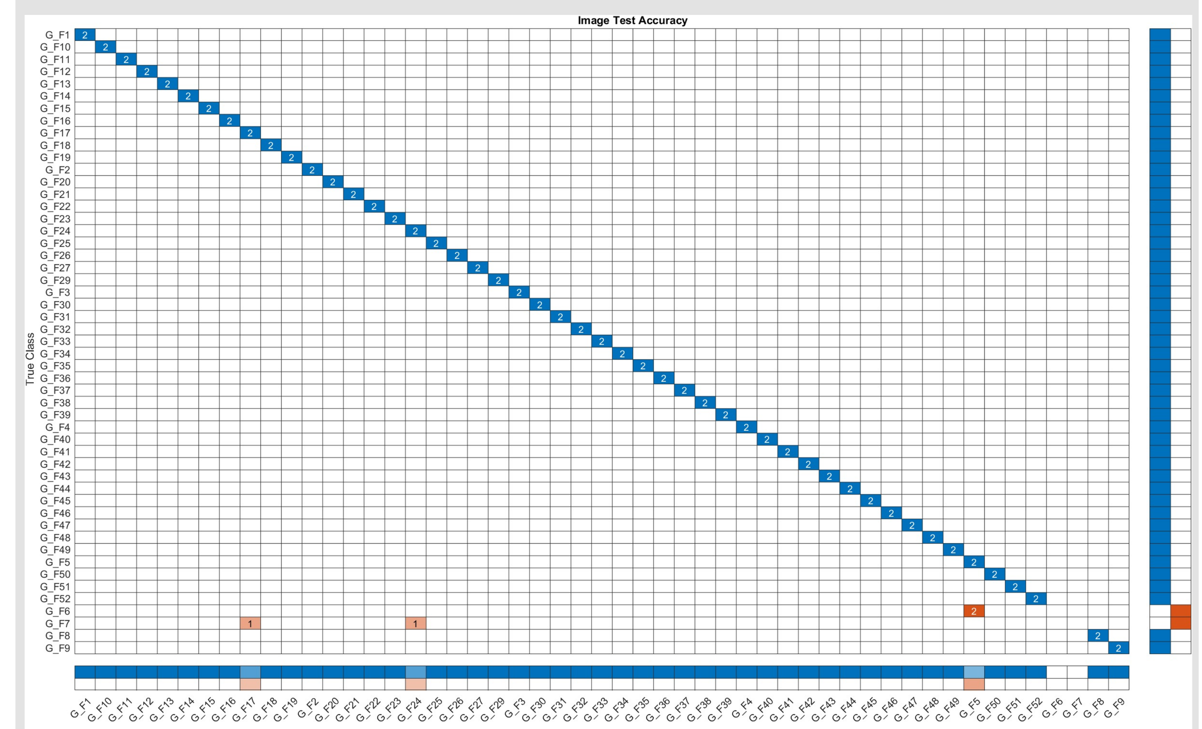


Figure 5. Confusion Matrix for the Face Recognition Test Results

Audio CNN:

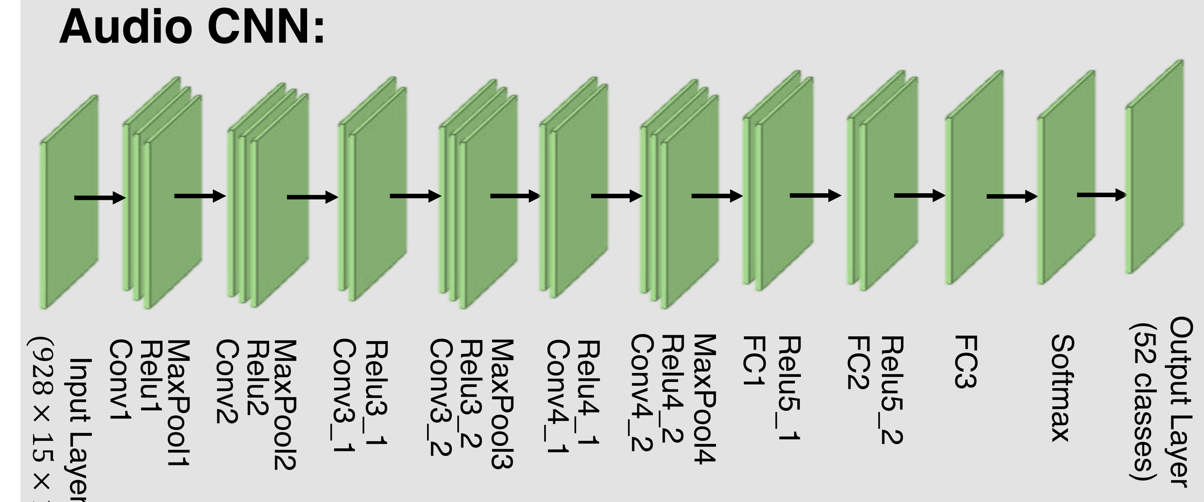


Figure 6. Layers of VGGish CNN after adjusting them to my problem

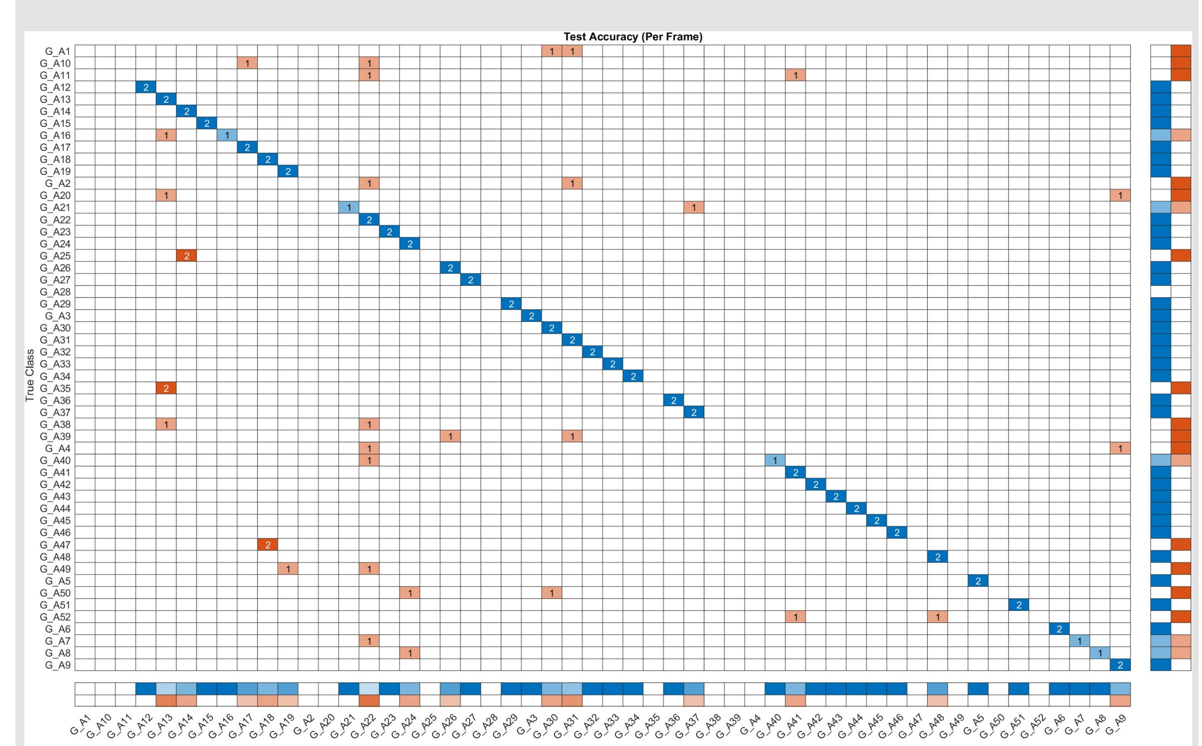


Figure 7. Confusion Matrix for the Soeaker Recognition Test Results

Fusion

Three decision-level fusion algorithms were implemented:

1. Fusing by higher confidence
2. Fusing by higher confidence after normalization
3. Fusing by higher entropy of confidence scores

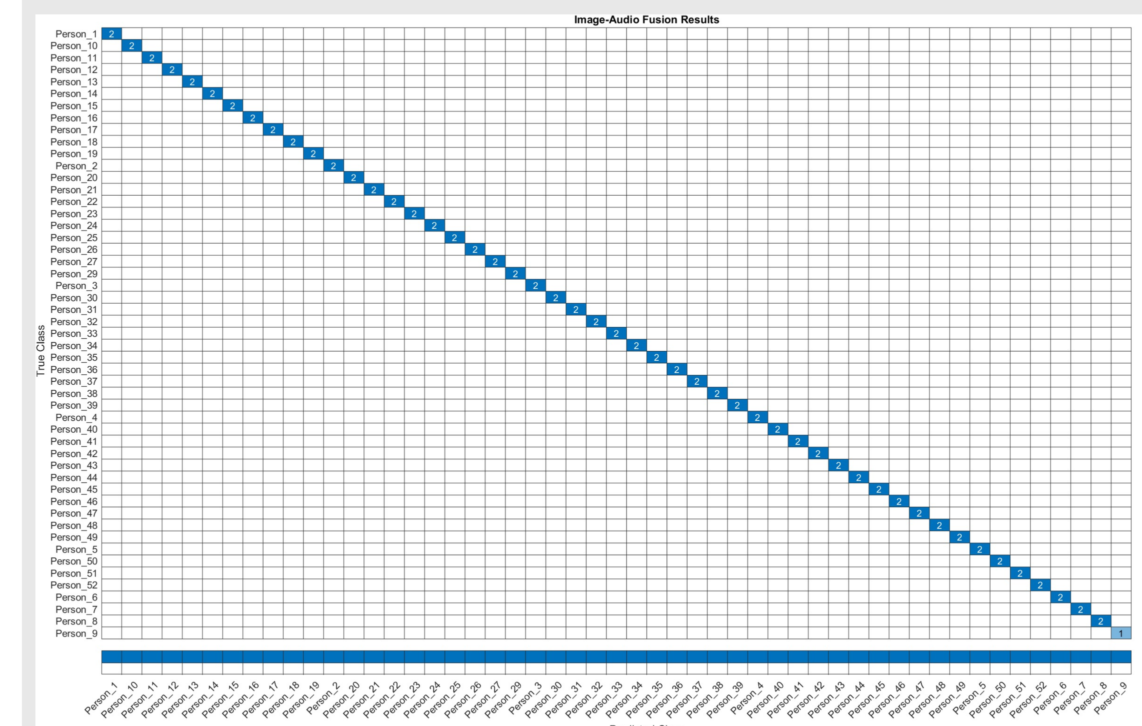


Figure 8. Confusion Marix for the Decision-Level Fusion Test Results

Conclusions & Future Directions

Results prove that biometric fusion improves the accuracy of person identification compared to using a single biometric.

Future work will include:

- Expanding the dataset to include more subjects
- Testing other convolutional neural networks
- Testing other fusion strategies (e.g., feature-level)
- Using greyscale edge detection for a color-blind face recognition system
- Implementing a two-stage system that first passes through a binary classifier in order to minimize the number of possible classes in the second stage

References

1. MSU Computer Vision Lab (n.d.). *MSU-AVIS Dataset* http://cvlab.cse.msu.edu/msu-avis-dataset.html
2. Wand, A (n.d.). *Facial Detection - Viola-Jones Algorithm* https://medium.com/@aaronward6210/facial-detection-understanding-viola-jones-algorithm-116d1a9db218
3. MATLAB (n.d.). *Speaker Identification using Pitch and MFCC* https://mathworks.com/help/audio/ug/speaker-identification-using-pitch-and-mfcc
4. Karimu, R (n.d.). *Mel-Frequency Cepstral Coefficients* https://apexpg.jimdofree.com/matlab-file/mel-frequency-cepstral-coefficients/

Acknowledgments

David and Lorraine Freed Undergraduate Research Symposium, Lehigh University