Project Number: 22
Project Title: Machine Learning Assisted Visual Speech Transcription
Project term: Spring 2019
Team members:
1. Safa Shaikh (POC) - safa.shaikh97@gmail.com
2. Timothy Liu - timmliu16888@gmail.com
3. Nick Lurski - nlurski@gmail.com
4. Avanish Mishra - avanish14@gmail.com

Project Advisor(s) name(s):
1. Dr. Kristin Dana

5. Up to 5 keywords that will help to classify the project scope:
   - Natural Language Processing
   - Computer Vision
   - Deep Learning
   - Mobile application
   - Client-server architecture

6. Project abstract (up to 250 words) to be shared with judges:

   In a world of multimedia and recorded events, understanding speech through video forms becomes relevant for all people. This project provides a platform for hearing-impaired persons to understand speech better by implementing programmed lip-reading. The current state-of-the-art is to extract audio information from videos and use the aural signature to categorize speech. Relying heavily on audio is a limiting factor in obtaining accurate video transcriptions due to noisy and unfamiliar speech, and we experiment with deep learning techniques to classify words spoken in the English language.
We teach a machine learning classifier to output the transcription of a received video. The classifier is trained with BBC news footage and utilizes computer vision to identify the lip movements. We use a ResNet architecture coupled with GRU layers and achieve over 80% accuracy in word identification. This classifier lives on a server that the user interacts with via an Android application. The application records and uploads video to the server, and displays the transcription back to the user. In addition, the application serves as a hearing-impaired toolkit by providing audio-to-text, text-to-speech, and a decibel meter. The server implements a content-aware compression system to make data transmission more efficient as well. This was implemented using a 2 dimensional DenseNet network, training on audiobook transcription.

Building a lexicon for our system to recognize speech is the first step in enabling continuous speech transcription. Coupled with audio, visual transcriptions can be more accurate and reliable, and when audio is unavailable, visual transcription becomes invaluable.