Speaker Identification for Voice Command-enabled Body Worn Cameras

Alex Hu, Ashwin Srinivas Badrinath, Christina Tang

Sponsor: Axon | Industry Mentors: Shwan Ashrafi, Ben Robaidek | Faculty Mentor: Mari Ostendorf

Problem Statement

Body worn cameras (BWCs) are used by law enforcement and police officers. In order to enable voice commands on BWCs, the BWC should only accept commands from verified officers.

In this system, an officer’s speaker features are saved onto the BWC during “enrollment.” Speaker features, from incoming raw audio, are compared to the stored speaker features in order to reject or accept a speaker.

Enrollment

1. Raw audio, or speech utterances, are converted to Mel-frequency cepstral coefficient (MFCC) vectors
2. Speaker embeddings are output from a convolutional neural network (CNN) and saved on BWC

Preprocessing

1. Train CNN as a classifier
   - VGG-M architecture
   - Cross-entropy loss function
2. Fine tune CNN with a Siamese network
   - Contrastive loss function
3. After, remove softmax
   - Get speaker embedding transform
   - Upload; no more retraining

End-to-End System

1. Get speaker embeddings from new speech utterances
2. Compare to saved embeddings with cosine scoring
3. Accept if average cosine score is above threshold

Experiment

Dataset
- 1251 total speakers
- 153516 total utterances
- Contains background noise

Training parameters
- CNN classifier
- Siamese fine-tuning
- Train 93% (1211 speakers)
- Validation 4% (1211 speakers)
- Test 3% (40 speakers)

Tools
- GTX 1070 Ti GPU
- TensorFlow
- Adadelta optimizer

Results

VGG-M architecture

- CNN EER = 33.78%
- Siamese EER = 34.62%

This system will make the wrong decision at an average of 33%. We were unable to replicate results shown in Siamese network fine-tuning research, because our system performed around the same.

Future Work

- Expand system to recognize a keyword
- Train on a dataset with speakers saying the keyword
- Add voice activity detection (VAD)
- Hands-free use

References