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

ELECTRICAL & COMPUTER ENGINEERING

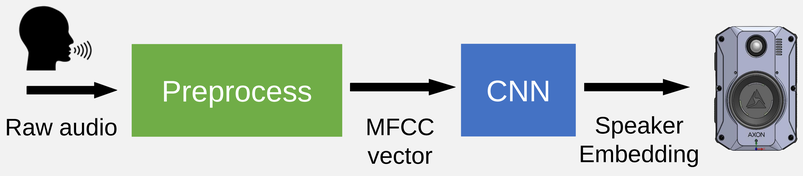
Problem Statement

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.

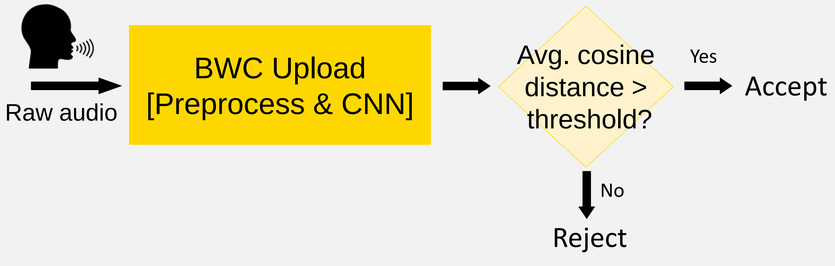
Enrollment

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



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



Preprocessing

Short-time Fourier Transform (STFT)
- 512 points
- 10 ms step size
- 25 ms Hamming window

Mapping
- 70 Hz to 8 KHz
- 40 bins in Mel-Spectrum

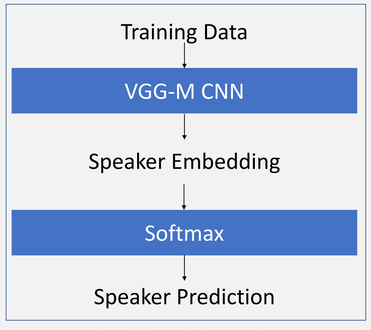
Discrete Cosine Transform (DCT)
- First 20 coefficients kept

CNN

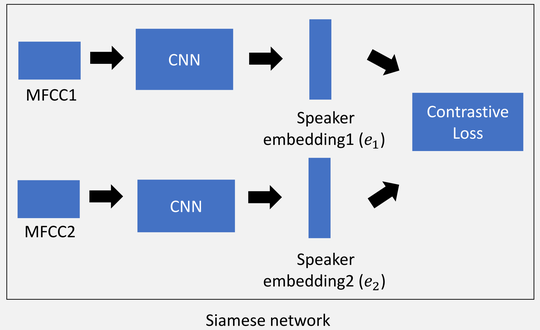
1. Train CNN as a classifier
   - VGG-M architecture
   - Cross-entropy loss function
2. Fine tune CNN with a Siamese network
   - Contrastive loss function
3. After, remove softmax
   - Get speaker embedding transform
   - Upload; no more retraining

Layer	Kernel	# Filters	Stride	Output size
Conv-1	7×7	32	2×2	$32 \times 17 \times 47$
Conv-2	5×5	64	1×1	$64 \times 13 \times 43$
Conv-3	3×3	128	1×1	$128 \times 11 \times 41$
Conv-4	3×3	256	1×1	$256 \times 9 \times 39$
Conv-5	3×3	256	1×1	$256 \times 7 \times 37$
fc-1	-	1024	-	-
fc-2	-	256	-	-
fc-3	-	1251	-	-

VGG-M architecture



Classifier training model



Siamese network

Experiment

Dataset

VoxCeleb1

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

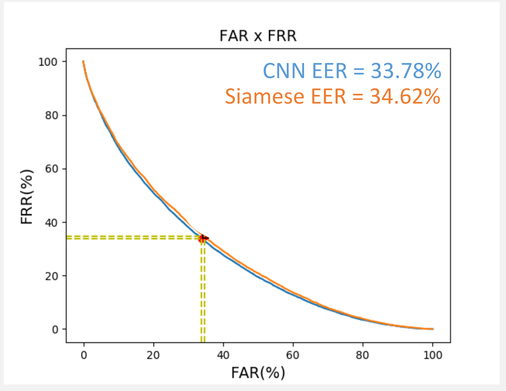
- Train 93% (1211 speakers)
- Validation 4% (1211 speakers)
- Test 3% (40 speakers)

Training parameters

	CNN classifier	Siamese fine-tuning
Epochs	20	20
Learning rate	.01	.0001
Batch size	128	128

Tools
- GTX 1070 Ti GPU
- TensorFlow
- Adadelda optimizer

Results



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

Apple (2018). *Personalized Hey Siri*.
Koch, G., Zemel, R. and Salakhutdinov, R. (2015). Siamese Neural Networks for One-shot Image Recognition. In: *32nd Conference on Machine Learning*. Journal of Machine Learning Research.
Nagrani, A., Chung, J. and Zisserman, A. (2017). *VoxCeleb: a large-scale speaker identification dataset*. Visual Geometry Group, University of Oxford.