Visual Speaker Identification in Noisy Environments

Gilbert Yap, Xinyue Zhou

BOSTON UNIVERSITY

Introduction

- Goal was to help identify if a person is speaking in a noisy environment using video data
- Audio-based speaker diarization relies on audio-based voice detection to identify speakers
- In videos, available data is a combination of visual and audio
  - Can use audio detection and video-based detection
- We compared Diarization Error Rate (DER) for both detection methods against a handmade annotation

Applications

- Speaker detection in video conferencing
  - Help highlight who is speaking, reduce false positives from background noise
  - Usable for any language
- Speech detection in audio-visual speaker diarization system
  - Can replace Voice Activity Detection (VAD) in noisy audio

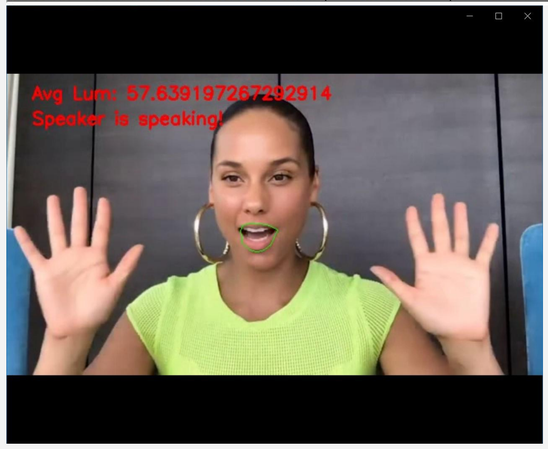
Methods

- Video Speaker Detection
  - Preprocess all frames first
  - Dlib facial landmarks to locate mouth
  - Get average luminosity of all pixels in mouth region
  - Threshold = Average Mouth Luminosity + Std Deviation
- Audio Detection
  - Python library “webrtcvad”
  - WebRTC implementation of voice detection

Results

- Terminology:
  - VSI - Visual Speaker Identification
  - VAD - Voice Activity Detection
- VSI method performed worse in noise-free environment compared to VAD method
- VSI method performed equal to or better than VAD method in ¾ cases where noise was added
- When the reference was trimmed to only include sections where a face is present, VSI method’s DER decreased about 10% on average

	VAD DER %	VSI DER %	VSI DER % (compared against only frames with face)
Noise-free Average	14.548	49.116	40.728
Audio + GWN Average	100	49.116	40.728
Audio + Piano Music Average	47.672	49.116	40.728
Audio + Jazz Music Average	50.622	49.116	40.728
Audio + Rain Sounds Average	80.348	49.116	40.728



Future Work

- Fine-tune luminosity calculation:
  - Add priority to center of mouth area
- Machine learning:
  - Implement method to track mouth luminosity for each face id
  - Use annotated dataset to differentiate smiling, bearded faces, etc. from speaking faces
- Increase dataset size
  - Make sure dataset is diverse
- Combine audio and video detection to decrease DER in audio-video file

References

Bredin, Herv'e ; Yin, Ruiqing ; Coria, Juan Manuel ; Gelly, Gregory ; Korshunov, Pavel ; Lavechin, Marvin ; Fustes, Diego ; Titeux, Hadrien ; Bouaziz, Wassim ; Gill, Marie-Philippe. pyannote.audio: neural building blocks for speaker diarization. IEEE International Conference on Acoustics, Speech, and Signal Processing. 2020.

Khan, Muhammad Usman Ghani; Mahmood, Sajid; Ahmed, Mahmood; Gotoh, Yoshihiko. Visual speech detection using OpenCV. Third International Conference on Open-Source Systems and Technologies. 2009.

Snyder, David; Chen, Guoguo; Povey, Daniel; MUSAN: A Music, Speech, and Noise Corpus. arXiv. 2015.

Ryant, Neville. dscore. GitHub. 2019.