Existing Computational Techniques for Audiovisual Data Analysis

(focusing on methods for data from classrooms)

T(CA)^2

The current state of automatic audiovisual analysis

This is still a rapidly advancing area, with current methods that can do things like:

- Measure synchrony and facial expressions
- Measure voice quality and conversational patterns

But large areas for improvement in classroom contexts remain:

- Analyzing videos with suboptimal lighting or camera position
- Extracting the contents of speech (the actual words)

What can audio data tell you about learning and classrooms?

- Body positioning
- Movement
- Gesture
- Hand raising
- Use of physical tools & materials
- Gaze
- Emotion

Types of videos

- Fixed cameras
  - Whole class with group-oriented seating
  - Whole class with front-facing students (mostly lectures)
  - Group-specific
  - Person-specific
  - Overhead (usually fisheye)
- Mobile cameras
  - Camera focused on one individual/group
  - Camera capturing one individual’s perspective
- And more!

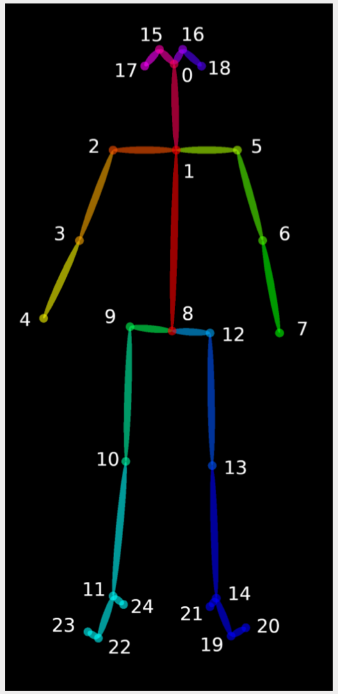
Whole class, group-oriented videos

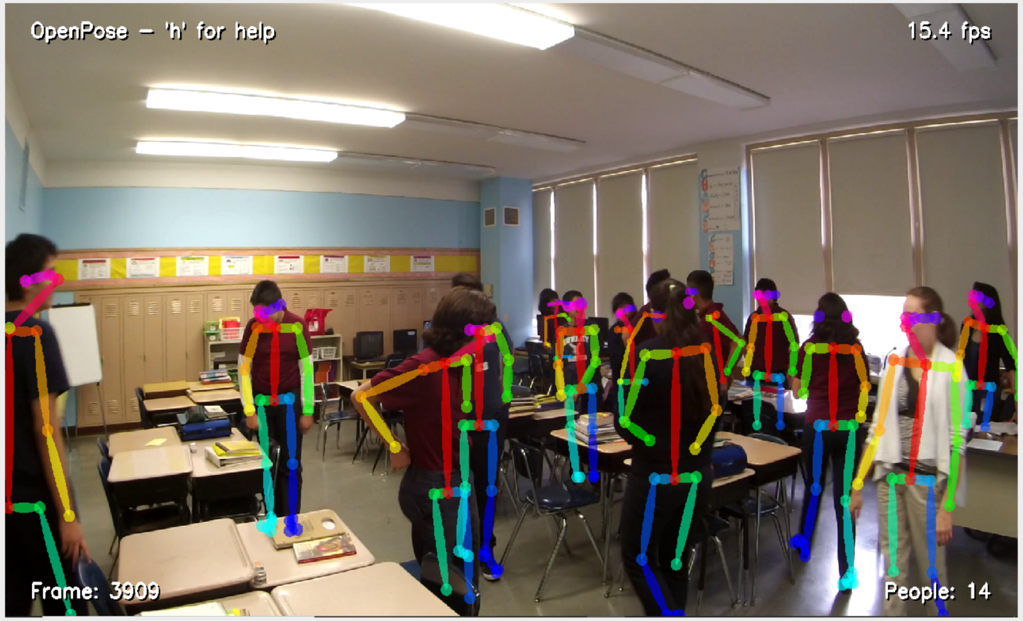
Characteristics:

- Teachers and students might move around
- Students often face each other
- Speech overlaps between students and teachers (no class-level turn taking)

OpenPose

- 25 key points:
    - Nose, neck, shoulder (R, L), elbow (R, L), Wrist (R, L), hip (R, M, L), knee (R, L), ankle (R, L), eye (R, L), ear (R, L), big toe (R, L), small toe (R, L), heel (R, L), background





T(CA)^2

Whole class lecture videos

Characteristics:

- Stadium seating or similar layout
- Clear distinction between student and teacher areas of the classroom
- Students typically face one way while the teacher faces them

EduSense

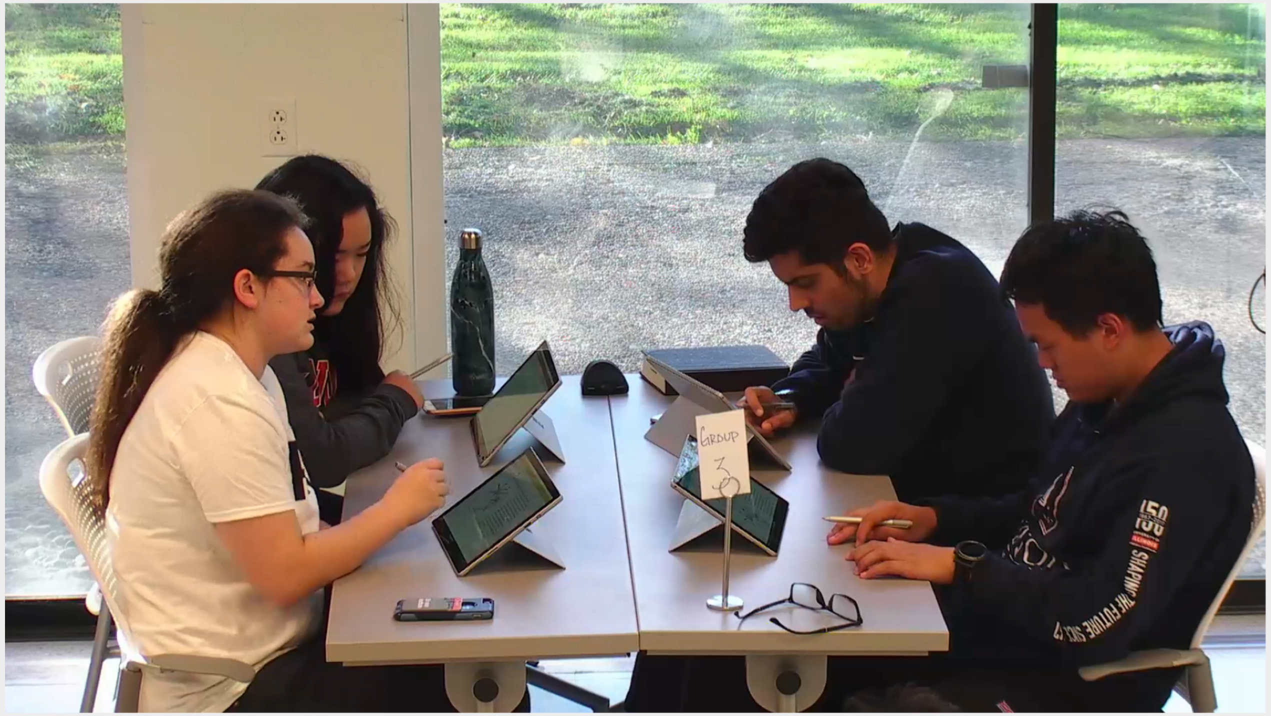
Extracts features from high-angle classroom videos, including things like:

- Hand raising
- Direction of gaze
- Sitting vs. standing

(also extracts instructor features with a teacher-facing instructor camera)

Group-specific videos

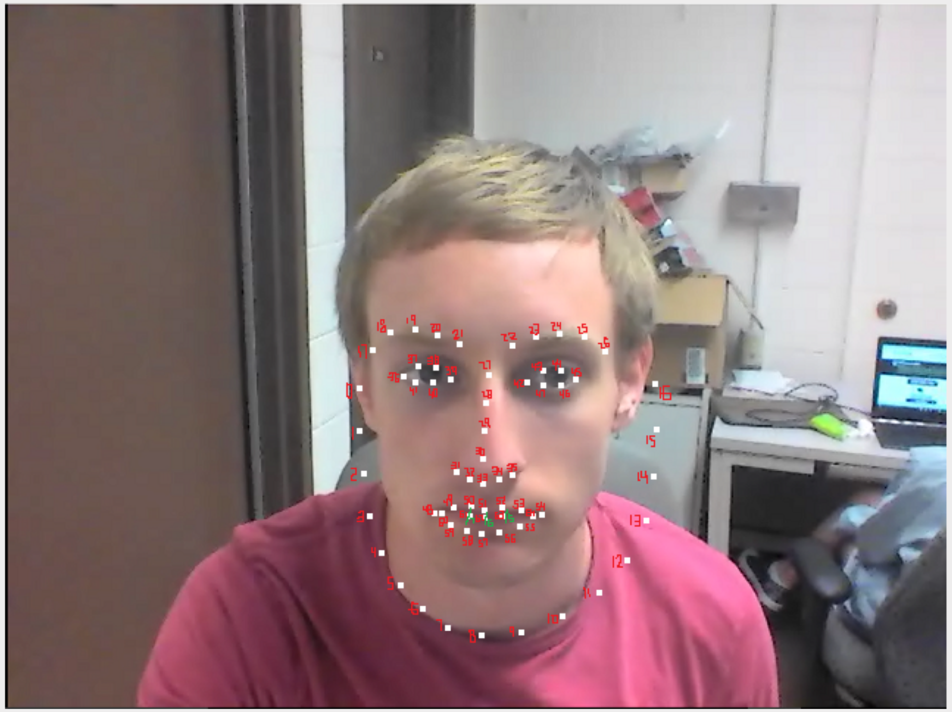
- Potential non-frontal views
- Instructor(s) may interlope



T(CA)²

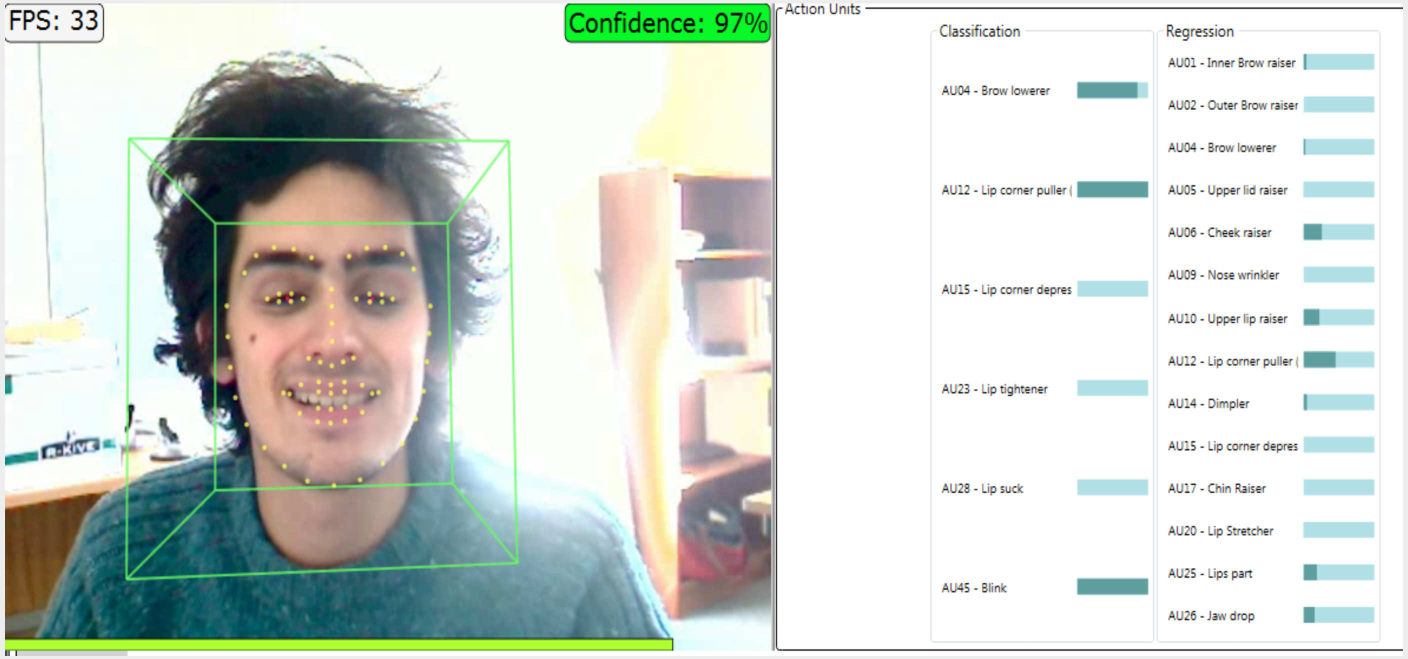
Person-specific videos

- Close-up and easily identifiable
- Limited to computerized or computer–mediated learning



OpenFace, FaceReader, and similar

- Action Units (AUs): Facial muscle movements
- High-level constructs like emotions (more context-dependent)



Which video types don’t have great solutions (yet)

- Fisheye camera lenses (e.g., overhead cameras)
- Mobile cameras (e.g., GoPro™)
- Videos where students or teachers are frequently occluded

What can audio data tell you about learning and classrooms?

- Participation share
- Discussion patterns
- Interaction with digital devices (like robots)
- Language learning
- Reading fluency
- Tutoring systems
- Assistive technologies for hard-of-hearing and visually impaired communities
- Early childhood contexts (using speech instead of typing for input)
- Collaboration and group work settings
- Question detection

$T(CA)^2$

Audio analysis terms

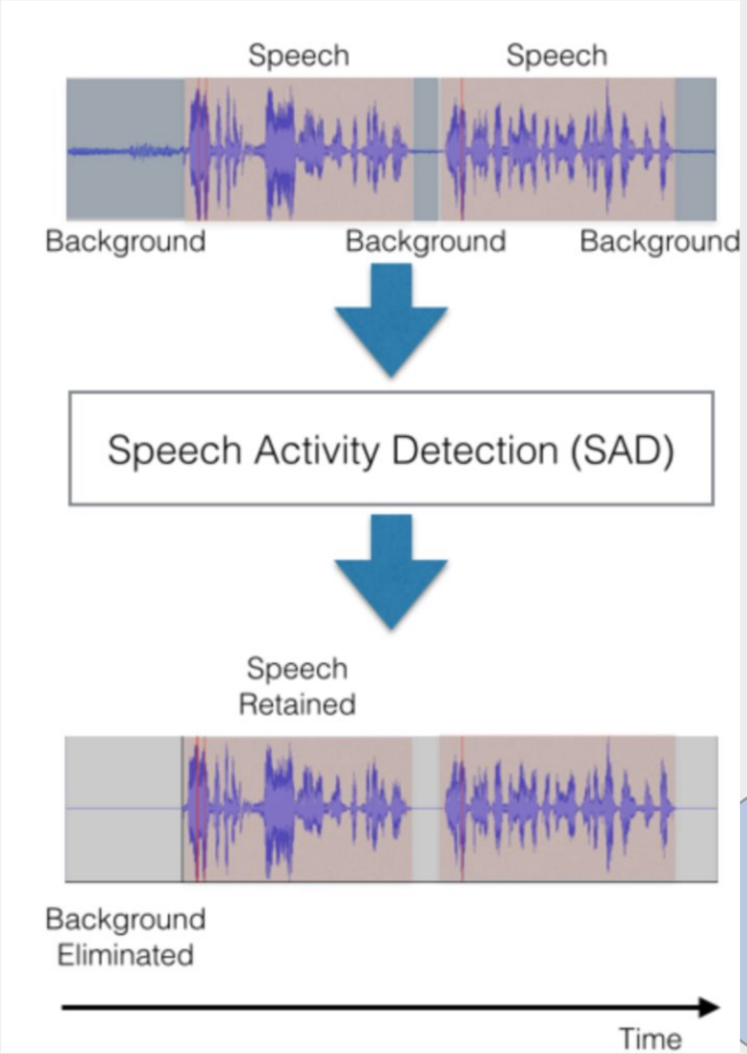
- SAD - speech activity detection
- ASR - automatic speech recognition
- prosody - qualities of speech (e.g., pitch, energy)
- speaker diarization - identifying individual speakers within a group (and separating them)
- NLP - natural language processing

Types of microphones

- Lapel mics
- Close-talking headset mics
- Table mics
- Microphone arrays
- Microphones embedded in phones/tablets/computers
- Special microphone devices (e.g., LENA)

Speech/Voice Activity Detection

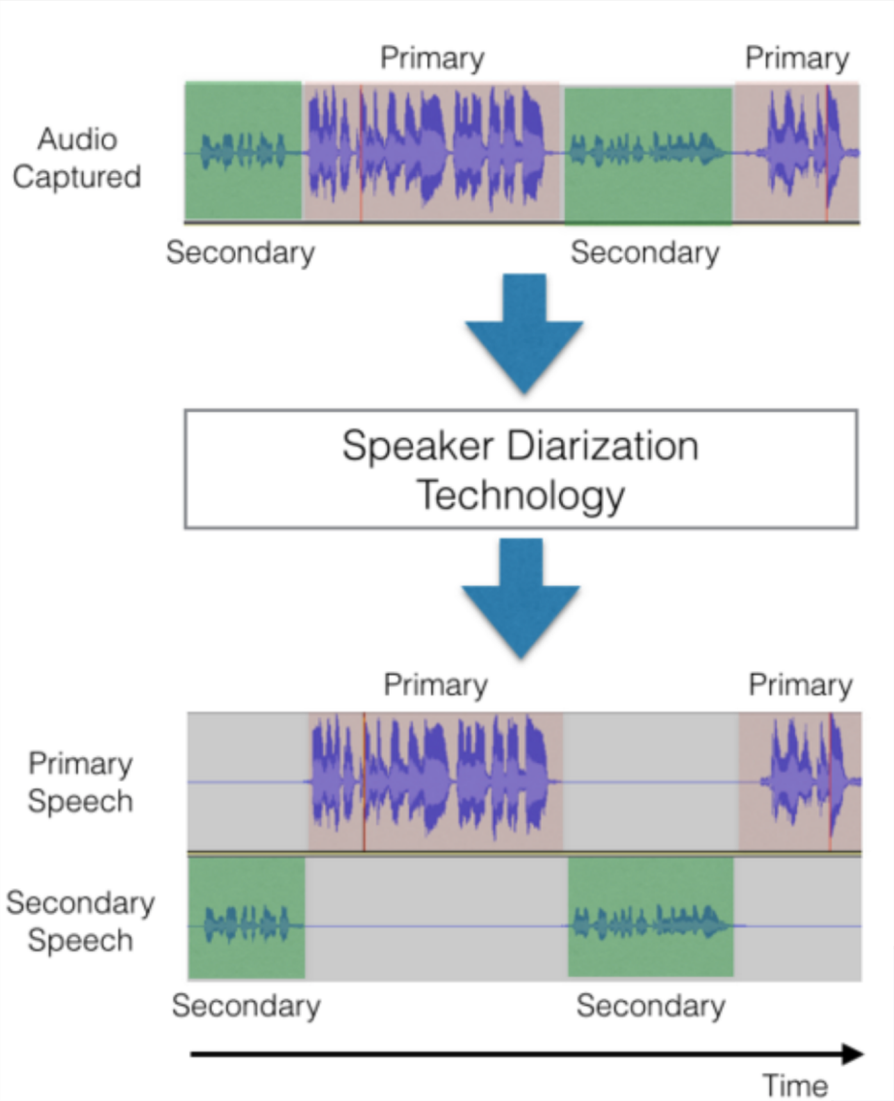
- Separates speech from acoustic background
- Can process large amounts of data efficiently; error rates around 10%



T(CA)²

Speaker Diarization

- Diarization algorithms help identity and separate individual speakers in a single audio track
- Error rates can be as low as 15% for naturalistic settings



T(CA)²

Audio analysis software options

- Praat
- HTK
- OpenEar
- Covarep
- ELAN
- ICSI Diarizer
- LIUM Diarizer
- CMU Sphinx
- Google ASR
- AT&T Watson ASR
- Bing Speech API
- Audacity
- Emovoice
- CMUSphinx
- WebRTC

Audio features

- Word Counts, Turn Counts, and Floor Sharing
- Sentiment detection (uses prosody)
- Detection of social signals (e.g., backchanneling)
- Laughter
- Filled vs. unfilled pauses
- Overlapped speech
- Stress detection

T(CA)^2

Audio analysis limitations

- Most work has been done on adults. Few data (corpora) of child speech exist
- Error rates still significantly high for Automated Speech Recognition of speech signals
- Youth speech development an issue
- Linguistic variation highly significant, both from youth to adults and across ages and regions
- Naturalistic speech patterns mostly unexplored
- Multiple-speaker interaction still a frontier
- Naturalistic acoustic environments challenging

T(CA)^2

Resources

- Video
  - OpenPose: https://github.com/CMU-Perceptual-Computing-Lab/openpose
  - OpenFace: https://github.com/TadasBaltrusaitis/OpenFace
  - FaceReader: https://www.noldus.com/facereader
  - EduSense: https://github.com/edusense/edusense

- Audio
  - Covarep: https://covarep.github.io/covarep/
  - Praat: https://www.fon.hum.uva.nl/praat/
  - OpenEar: https://github.com/moneriomaa/openear