ACTIVE SPEAKER DETECTION

How Technology Hears the Loudest Voice

PROJECT INTRODUCTION

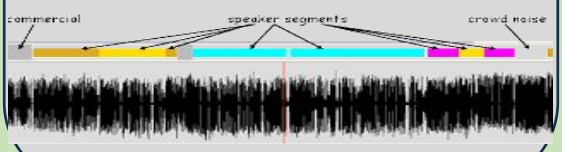
- Enhanced accuracy through audio-visual fusion.
- Unified model outputs for precise speaker detection
- Innovative solution for dynamic multi-speaker recognition.

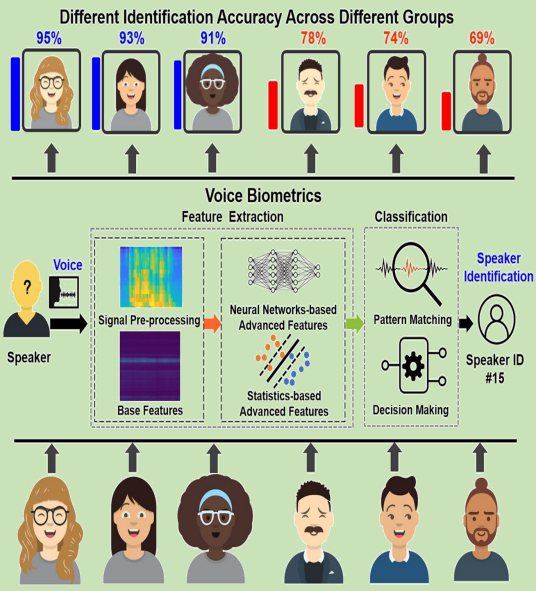
METHODOLOGY

- Fusion of audio and visual data enhances accurate speaker detection.
- Audio model: CNN for speaker classification, GMM for change detection.
- Video model: Frame-wise analysis, Retina Face for face detection.

AUDIO MODEL

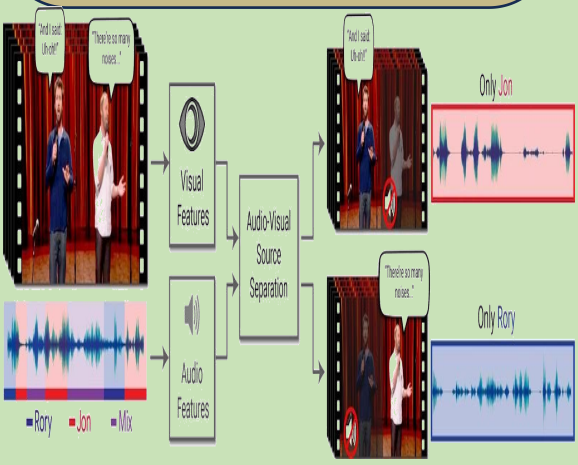
- Utilized Convolutional Neural Network (CNN) for audio feature extraction and classification.
- Implemented multi-class classification with 4 labels
- Final prediction indicates the type of audio content present.





SPEAKER CHANGE MODEL

- Employed Voice Activity Detection to identify segments with speech in the audio.
- Employed GMM-based clustering to discern speaker boundaries.
- Achieved accurate identification of transitions between different speakers.



AV FUSION

- Fusion of all the model' s score.
- Find max active speaker score.
- Pan the camera to the active predicted speaker based on the max score

FUTURE WORK

- Integrate emotion detection for added contextual insights.
- Develop a user-friendly real-time interface for easy visualization.

VIDEO MODEL

- Utilizing Retina Face for accurate face detection and alignment.
- Overcoming difficulties in video contexts with robust face localization.

PREPARED BY:
KASHAF KHAN:CS-19002
HAFSA ZAFAR:CS-19012
ALEESHA AHMED:CS-19013
INTERNAL:DR.UROOJ AINUDDIN
(ASSISTANT PROFESSOR)

DEPARTMENT OF COMPUTER AND INFORMATION SYSTEMS ENGINEERING