EyeHear

A vision-audio fusion system for robust speech recognition

Team D6
Larry Geng, Stella Getz, Chao Li (Charlie)

Content

- Motivation
- Use Case
- Requirements for Use Case
- Technical Challenges
- Solution Approach
- Design Challenges and Mitigations
- Testing Setup and Verification Metrics
- Tasks and Division of Labour
- Schedule
- Conclusion

Use Case

Speech-to-Text

Mixed Speech

OR RATHER OF HIS DANGERLE RAILS
SILLOMS DEAL LIKE THA ATER OER

Problem

Modern visual and/or audio system cannot *easily* distinguish speech from multiple speakers

EyeHear

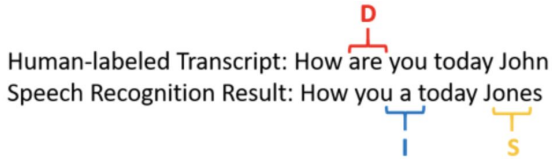
EyeHear produces an ***enhanced real-time video with captions*** for each speaker

Areas

Software, Hardware and Signal Systems

Requirements for Use Case

1. Accuracy of NLP output:
   - **Word Error Rate (WER)** <= 20%
2. Quality of video played for user
   - Video: Suitable for live streaming, 720p (HD)
3. Quality of audio collected by mics
   - Audio: Voice between 20Hz and 4kHz
4. Delay: Capture device to display
   - < 150ms recommended by Zoom
5. Delay: Capture device to captions
   - 2s delay on Zoom live captions
6. Physical size of device:
   - Can sit on the end of an average conference table
   - < 4.02lb (weight of Macbook Pro)

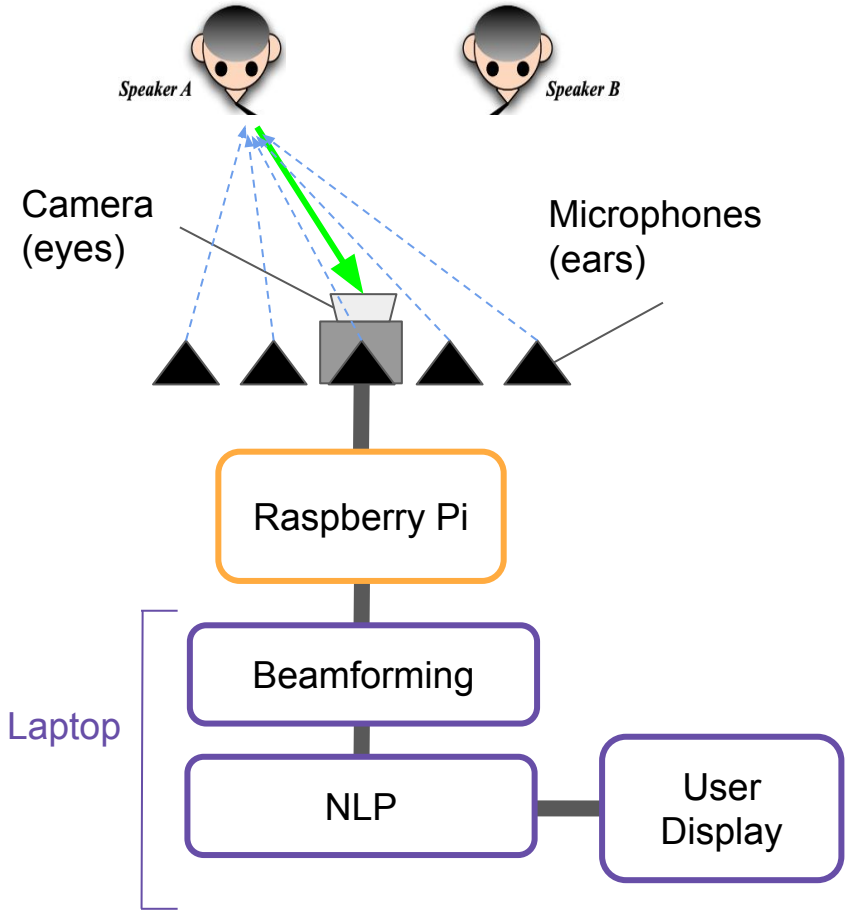


$$WER = \frac{I + D + S}{N} * 100\%$$

Technical Challenges

1. Accurate speech separation and NLP output
2. Low latency processing pipeline
3. High quality video and audio
4. Portability of device

Solution Approach

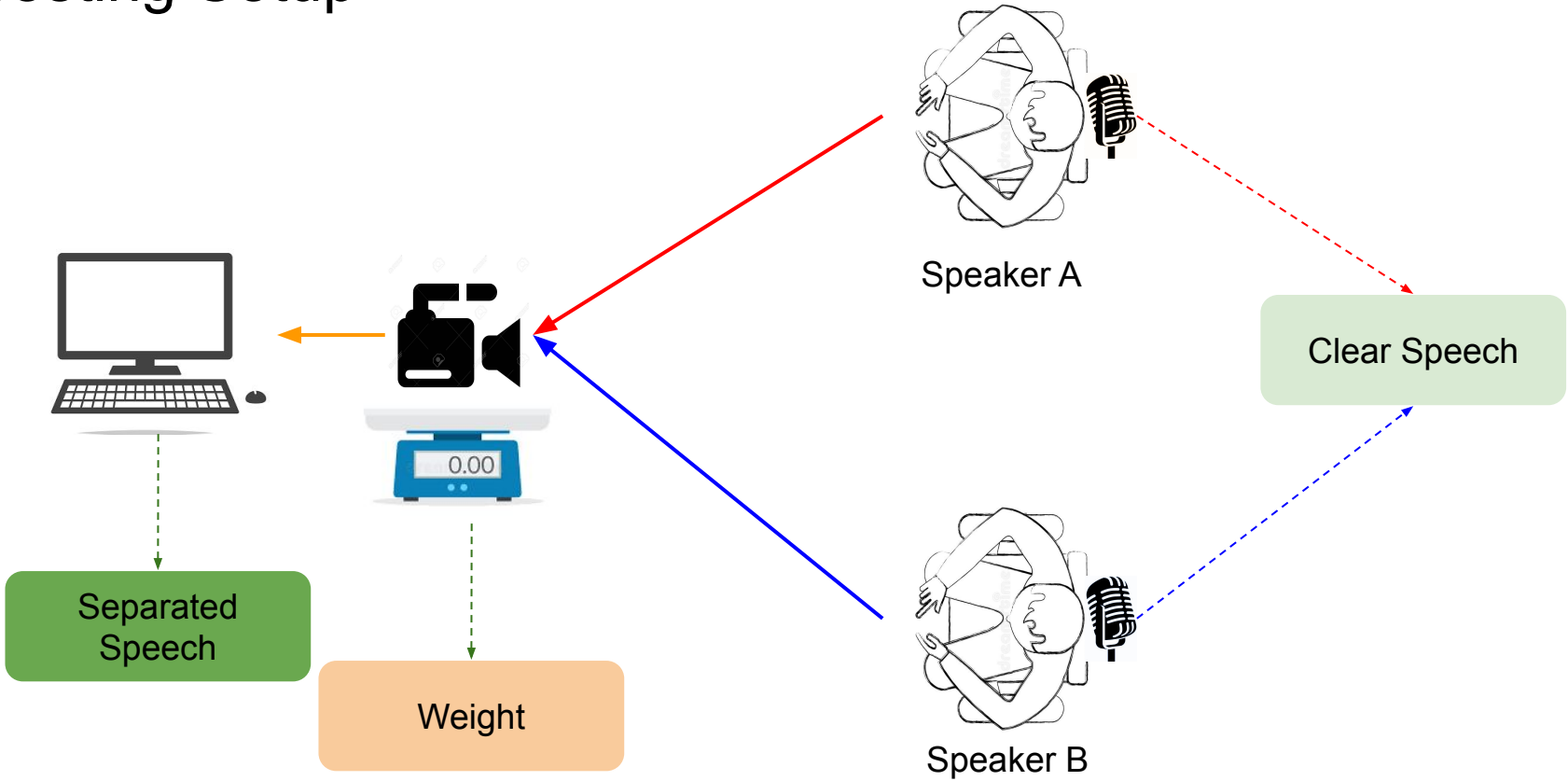


1. Accurate speech separation and NLP output
   - Beamforming with image segmentation
   - Using IBM Watson Speech-to-Text (~5% WER)
2. Low latency processing pipeline
   - Single calibration step to determine speaker location
3. High quality video and audio
   - 30fps Camera
   - 8kHz Microphone sampling
4. Portability of device:
   - Fully enclosed device
   - 6 microphones of 7.1 ounces each

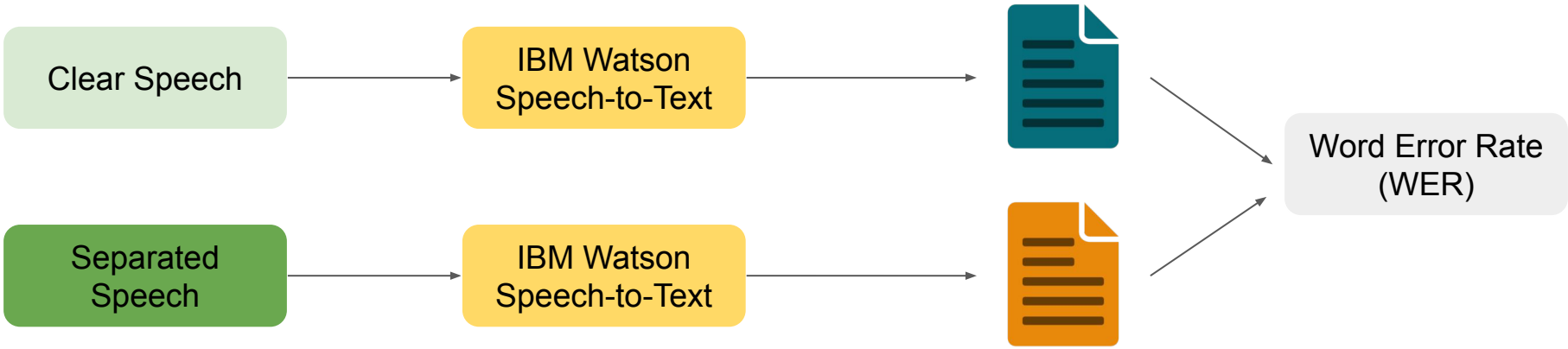
Design Challenges and Mitigations

Challenge	Mitigation
Beamforming does not allow us to sufficiently enhance target speech	Use Short-Time-Fourier-Transform (STFT) methods to separate speech instead
Unable to synchronize or sufficiently sample from microphone array	Use two single-channel pre-built microphone arrays
Inaccurate camera estimation of the location of speakers	Averaging angle estimate from camera with angle estimate from microphone array
Laptop cannot complete processing in little enough time for real-time captioning	Record video and transcribe scripts rather than playing a captioned video in real time for user

Testing Setup



Verification Metrics



Tests
Compare **WER** of filtered isolated speech and mixed speech
Compare **WER** of filtered isolated speech and clear isolated speech

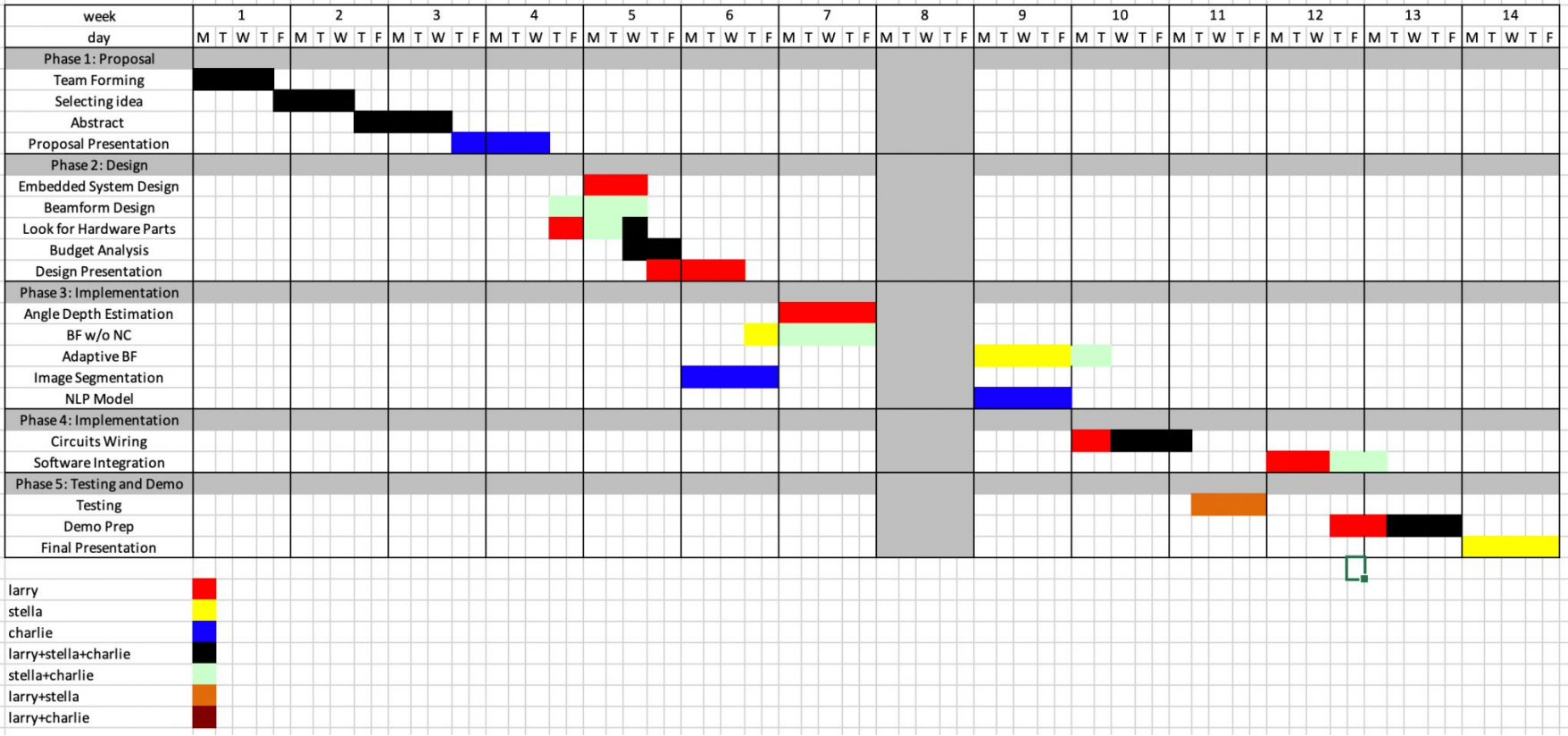
Verification Metrics

Requirement	Test
Quality of video played for user: Suitable for live streaming, 720p	Record video and check that output is 720p
Quality of audio collected by mics: Voice, 20Hz to 8kHz sampling	Play sine sweep from 20Hz to 8kHz and plot played vs. recorded audio sweep of frequencies
Delay: Capture device to display: <150ms recommended by Zoom	Using a 60fps phone camera, record video/audio input and video/audio output to measure latency
Delay: Capture device to captions: <2s delay on Zoom live captions	Using a 60fps phone camera, record speaker and display. Calculate time between the first frame of a spoken word and the first frame the caption appears in.

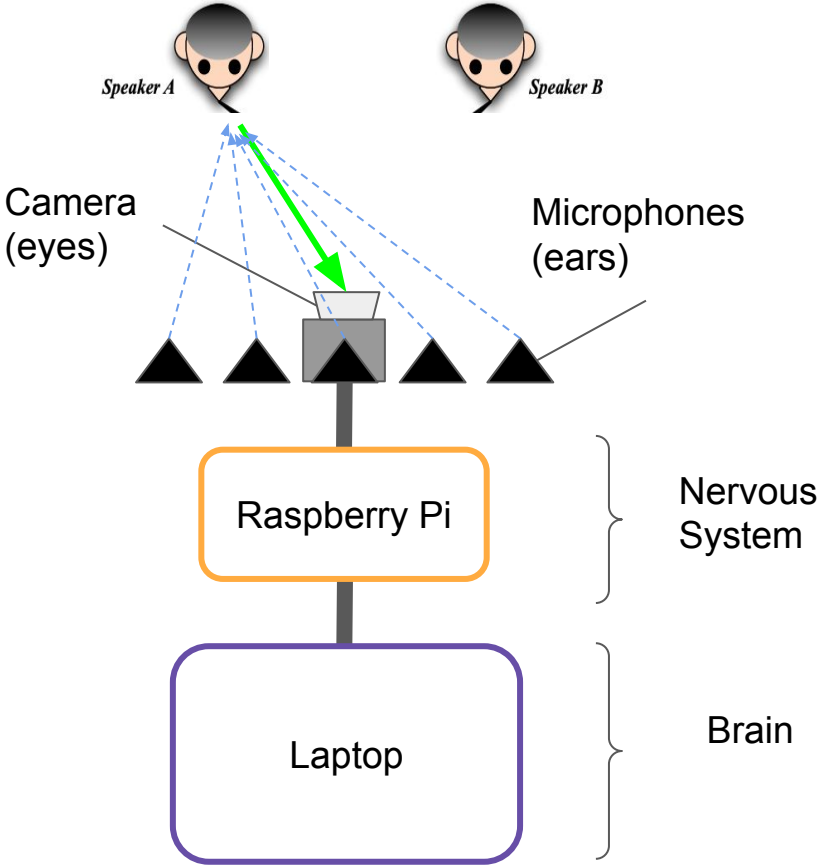
Tasks and Division of Labor

- Pre-built
  - Natural Language Processing (NLP) software
  - Image segmentation software
- Larry Geng
  - Image Processing
  - Device Setup
- Stella Getz
  - Image Processing
  - Audio Processing
- Charlie
  - Deep Learning (Image Segmentation, Natural Language Processing)
  - Audio Processing

Schedule


week
1 2 3 4 5 6 7 8 9 10 11 12 13 14
day
M T W T F
Phase 1: Proposal
Team Forming
Selecting idea
Abstract
Proposal Presentation
Phase 2: Design
Embedded System Design
Beamform Design
Look for Hardware Parts
Budget Analysis
Design Presentation
Phase 3: Implementation
Angle Depth Estimation
BF w/o NC
Adaptive BF
Image Segmentation
NLP Model
Phase 4: Implementation
Circuits Wiring
Software Integration
Phase 5: Testing and Demo
Testing
Demo Prep
Final Presentation
larry
stella
charlie
larry+stella+charlie
stella+charlie
larry+stella
larry+charlie

Conclusion



Stephen Colbert Show
17 August 2021