Team D6 - EyeHear

Stella Getz, Larry Geng, Charlie Li

Add your 12 slides after this slide… [remember, 12 min talk + 3 min Q/A]

For more information about formatting or importing slides see:
https://gsuite.google.com/learning-center/products/slides/get-started/

Make sure to cover
(refer to the Final Presentation Guidance):

- Use Case / Application and Primary (Quantitative) Requirements (i.e. A reminder from prior presentations)
- Solution Approach – a reminder (include updates from Design Review presentation if changed)
- System Implementation – your complete solution
- Testing, Verification and Validation – with quantitative Metrics and target values to compare with experiment
  - What tests did you run ? How many tests ? What were the results ?
  - Graphs, tables, quantitative results (compare with the metric targets & ultimately use-case requirements)
- Project Management – Tasks, division of labor, and schedule
- Lessons Learned

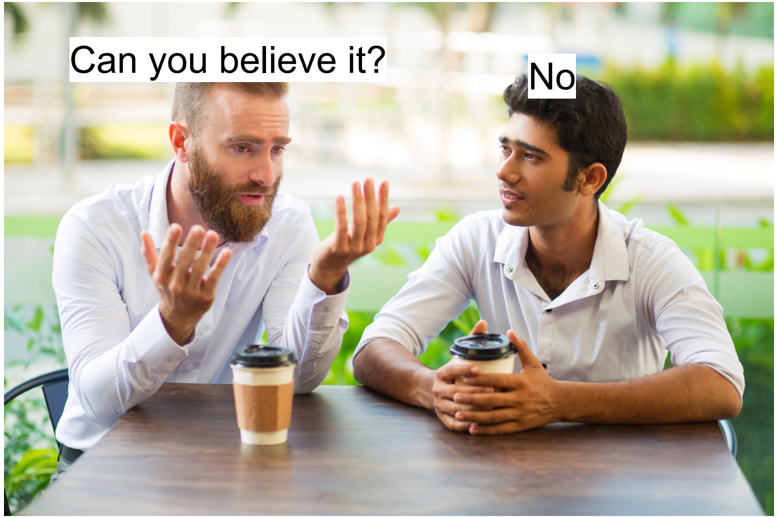
Consider that this slide already works as a introduction slide so use your first slide wisely

Use Case

Real-time video with captions for each speaker.

Visually match speakers with what they say.

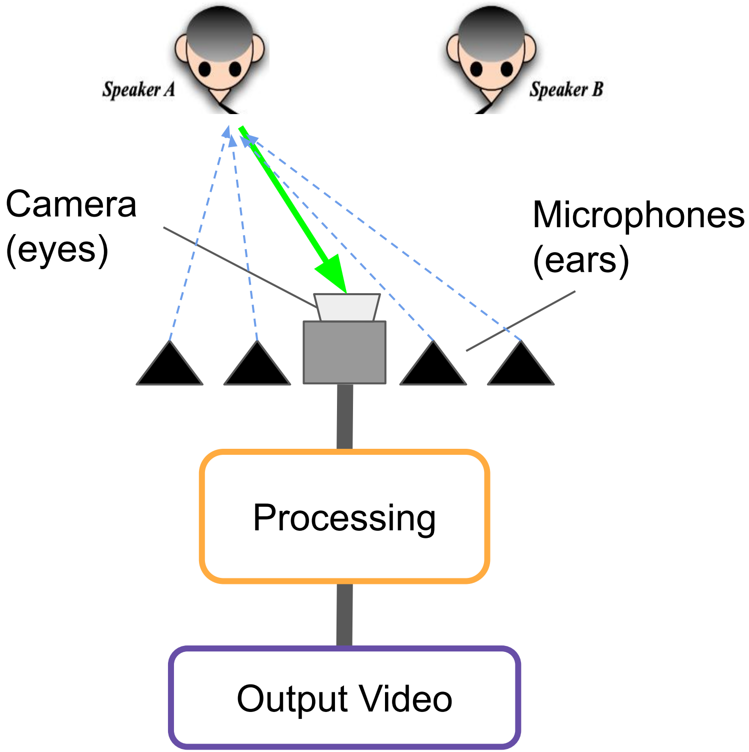
Enhance live meetings and recordings with better captions.



Use Case Requirements

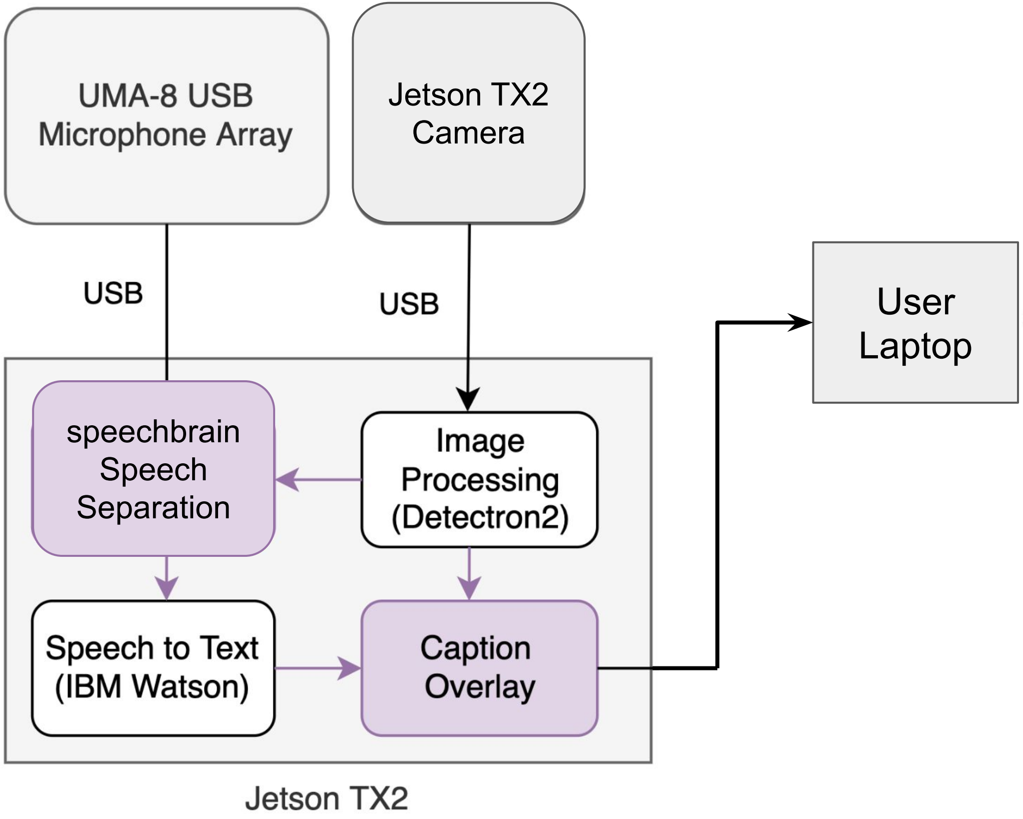
Metric	**Requirement**	**What we achieved**
Accuracy of speech to text output	Word Error Rate <= 20%, Microsoft standard	[results discussed later]
Minimum number of speakers	2	Achieved
Delay: Video+audio capture to captioned video display	< 2s delay, Zoom live captions	No longer pursuing real-time: Requirement changed to < 1min processing time (Achieved)
Quality of video played for user	720p (HD) 30fps, suitable for live streaming	Achieved
Range of audio capture by microphones	75 Hz to 4kHz, range of human voice	Achieved
Size and weight of device	Can fit on conference table, < 4.0lb	Can fit comfortably on conference table (photo later)

Final Solution Approach

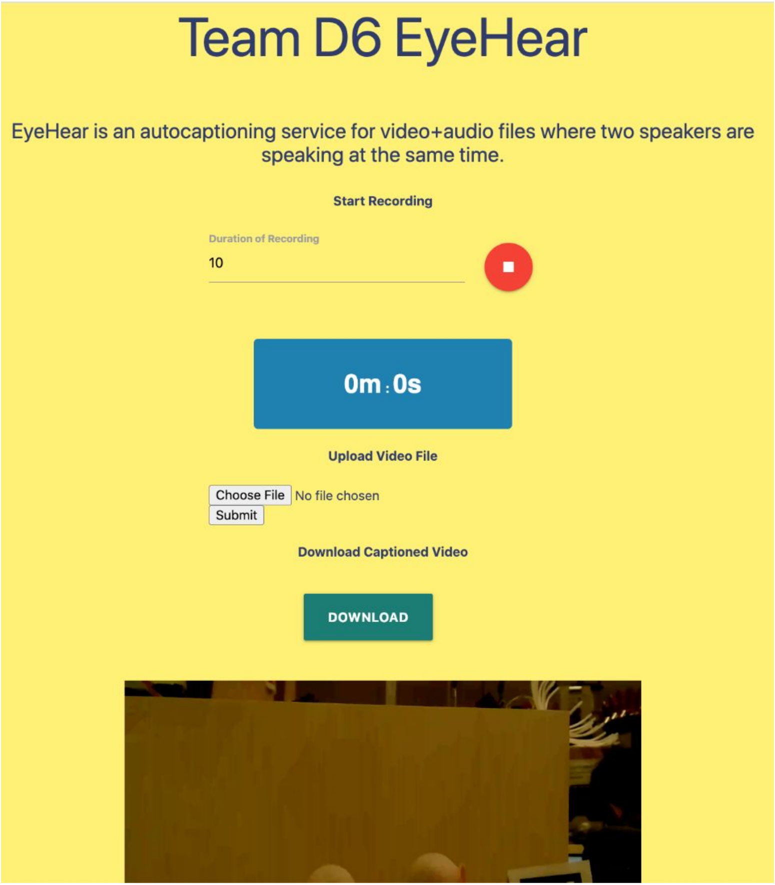


1. Camera identifies location of speakers for positioning of captions
2. Deep learning algorithm uses stereo audio recording to separate speakers
3. Beamforming concepts help determine which speaker is left and which is right
4. Audio is fed to NLP model to produce captions
5. Captions overlaid over final video

System Diagram


UMA-8 USB Microphone Array
Jetson TX2 Camera
USB
USB
User Laptop
speechbrain Speech Separation
Image Processing (Detectron2)
Speech to Text (IBM Watson)
Caption Overlay
Jetson TX2

Website:



Team D6 EyeHear

EyeHear is an autocaptioning service for video+audio files where two speakers are speaking at the same time.

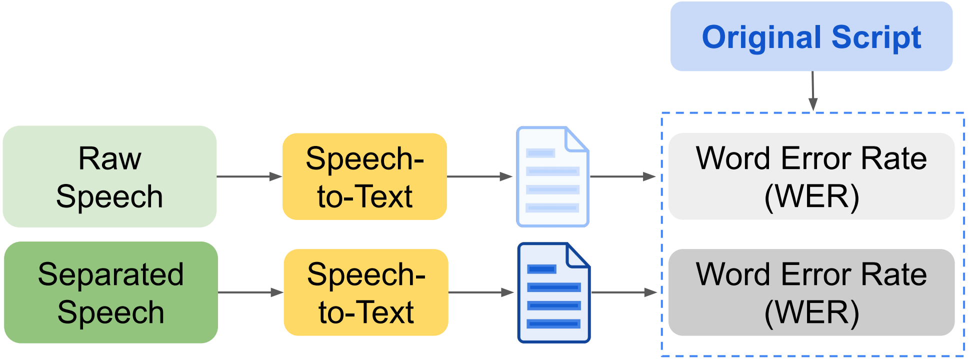
Start Recording

Duration of Recording: 10

Upload Video File

Download Captioned Video

Testing, Verification, and Validation

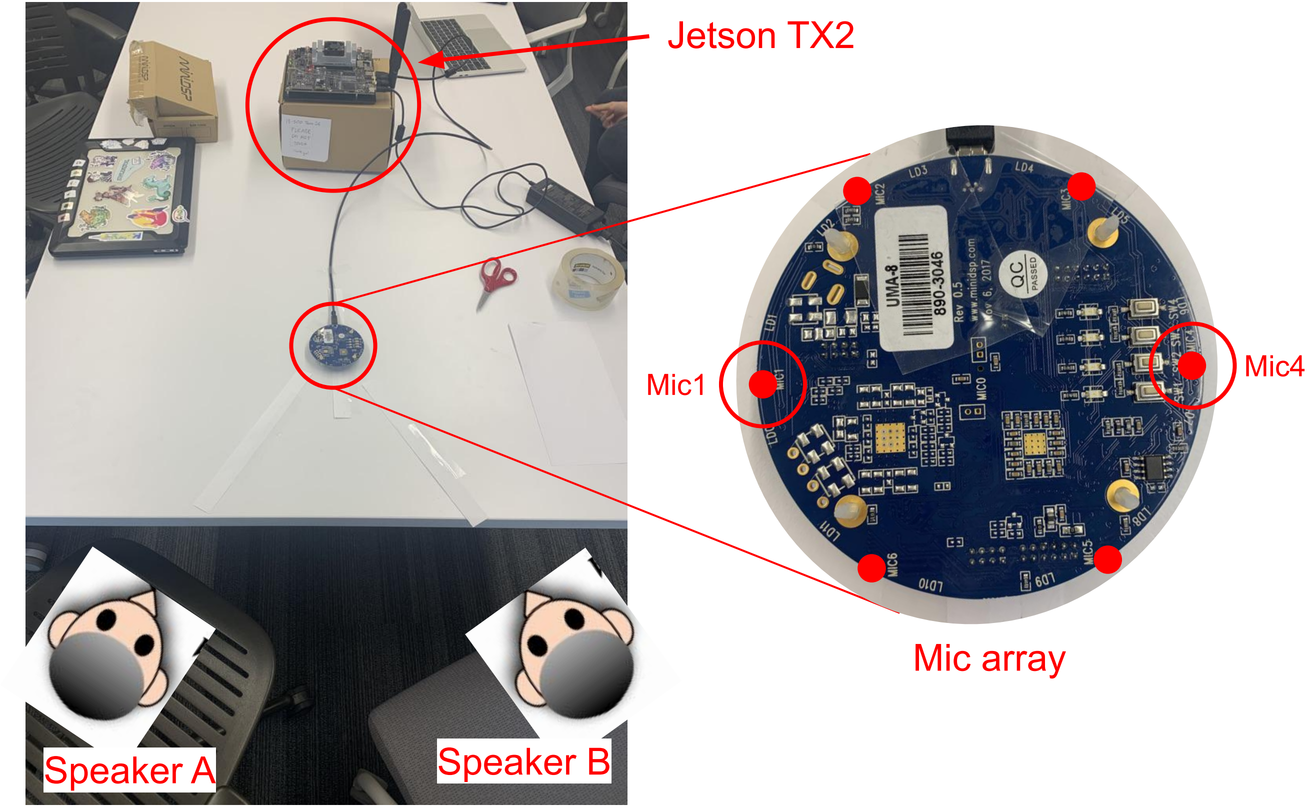
Metric	**Target Value**	**Test**
Accuracy of speech to text output	Word Error Rate (WER) <= 20%	 Original Script Raw Speech Speech-to-Text Word Error Rate (WER) Separated Speech Speech-to-Text Word Error Rate (WER)
Minimum number of speakers	Acceptable WER for 2 speakers	Measure WER degradation when adding additional speakers.
Quality of video played for user	1280 x 720 pixels	On laptop end, check output resolution and fps
Processing time	< 1min	Measure time from end of recording to when the captioned video is ready

Speech Separation: Signal Processing vs. Deep Learning

Raw Audio		SSF + PDCW		Deep Learning Approach	
Speaker Left WER	Speaker Right WER	Speaker Left WER	Speaker Right WER	Speaker Left WER	Speaker Right WER
97.1	127.3	-	66.7	41.2	54.5

→ Conclusion: Proceed with deep learning

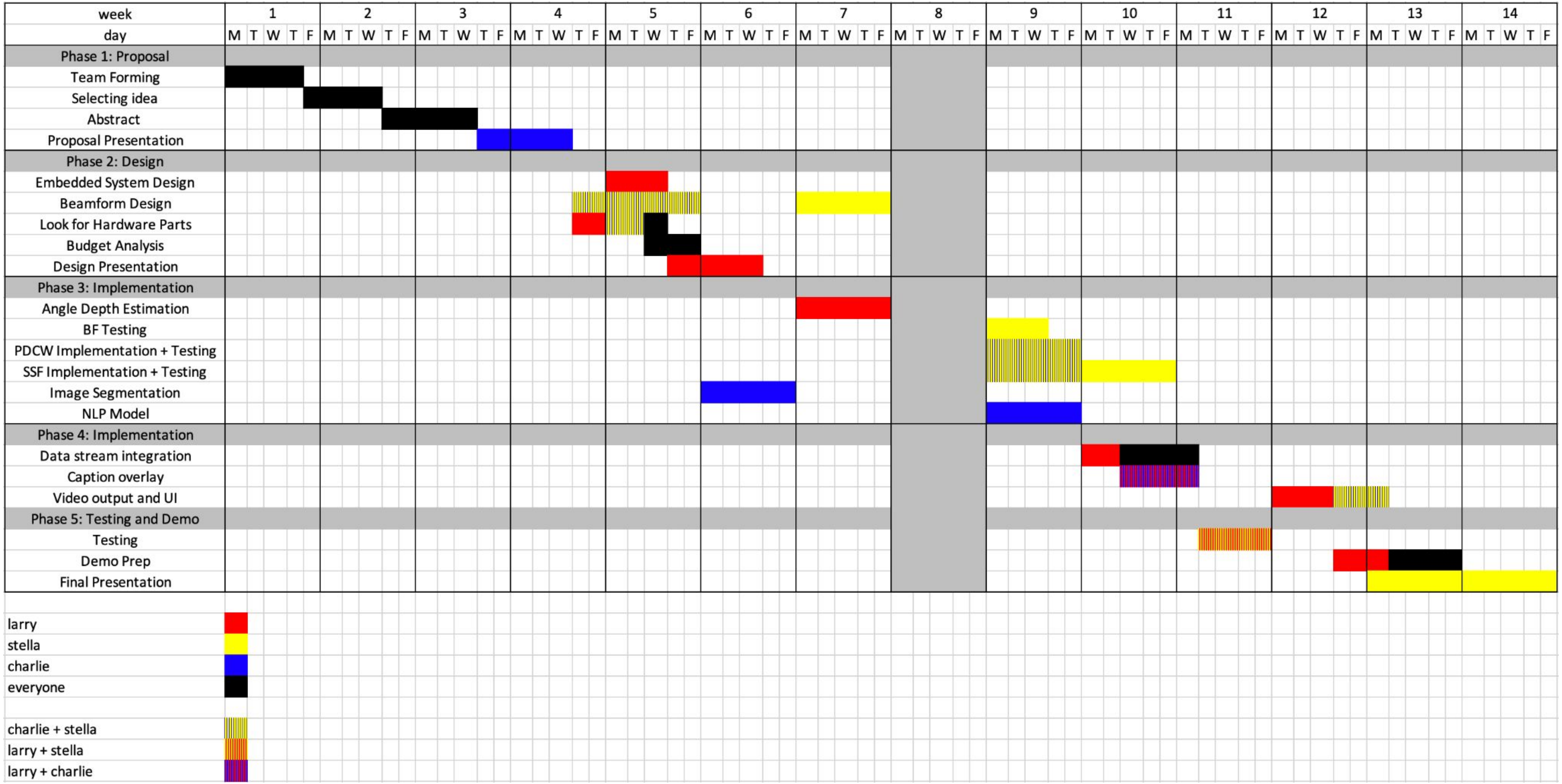
System Implementation & Testing Setup



Testing Results - Accuracy of speech to text output

	Raw Signal				**Deep Learning Approach**			
	conference room		CUC loggia		conference room		CUC loggia	
Metric	Speaker Left WER	Speaker Right WER	Speaker Left WER	Speaker Right WER	Speaker Left WER	Speaker Right WER	Speaker Left WER	Speaker Right WER
Speaker Left	33.3	-	29.6	-	6.7	-	23.3	-
Speaker Right	-	39.4	-	84.4	-	43.2	-	67.6
Speaker Right interrupted by Speaker Left	-	87.5	-	71.8	-	40.5	-	59.5
Speaker Left partially overlapped by Speaker Right	62.2	84.4	65.2	87.0	36.7	51.4	36.7	54.1
Speaker Left completely overlapped by Speaker Right	93.8	109.4	97.7	88.6	43.3	73.0	70.0	73.0

Schedule and Work Distribution



week	1	2	3	4	5	6	7	8	9	10	11	12	13	14
day	M T W T F	M T W T F	M T W T F	M T W T F	M T W T F	M T W T F	M T W T F	M T W T F	M T W T F	M T W T F	M T W T F	M T W T F	M T W T F	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 Testing									🟨🟨🟨··					
PDCW Implementation + Testing									▥▥▥▥▥					
SSF Implementation + Testing									▥▥▥▥▥	🟨🟨🟨🟨🟨				
Image Segmentation						🟦🟦🟦🟦🟦								
NLP Model									🟦🟦🟦🟦🟦					
Phase 4: Implementation														
Data stream integration										🟥🟥⬛⬛⬛	⬛····			
Caption overlay										··▦▦▦	▦····			
Video output and UI												🟥🟥🟥▥▥	▥····	
Phase 5: Testing and Demo														
Testing											·▤▤▤▤			
Demo Prep												···🟥🟥	🟥⬛⬛⬛⬛	
Final Presentation													🟨🟨🟨🟨🟨	🟨🟨🟨🟨🟨

larry	🟥
stella	🟨
charlie	🟦
everyone	⬛
charlie + stella	▥
larry + stella	▤
larry + charlie	▦

Final Product

Next Steps

- Try SSF + PDCW pre-processing
- Calculate new WERs when using better mic setup