D6: EyeHear

Larry Geng, Stella Getz, Chao Li (Charlie)

Content

- Use case
- Solution approach
- Block diagram and implementation plan
- Planned testing
- Risks and mitigations
- Schedule and division of labor

Use Case

Real-time video with captions for each speaker.

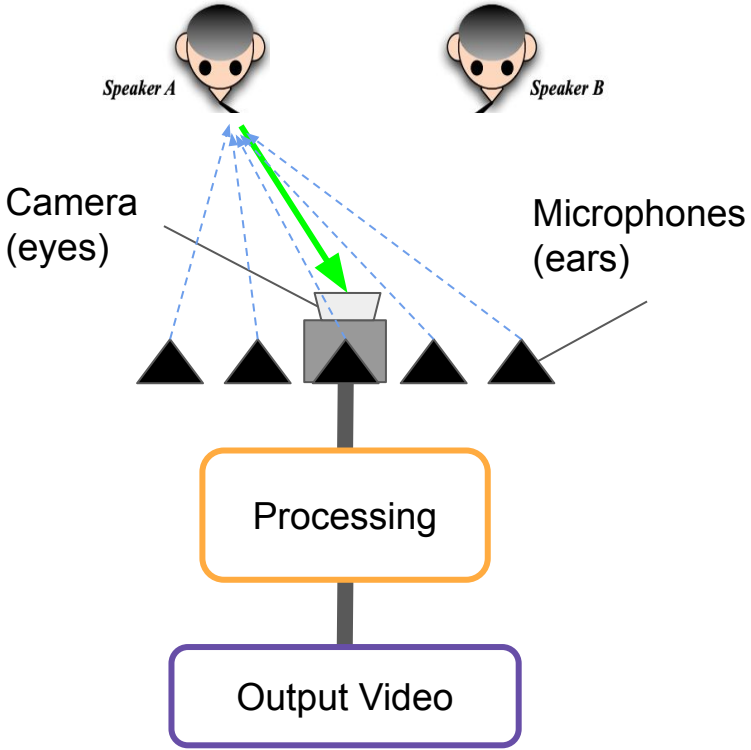
Visually match speakers with what they say.

Enhance live meetings and recordings with better captions.

Use Case Requirements

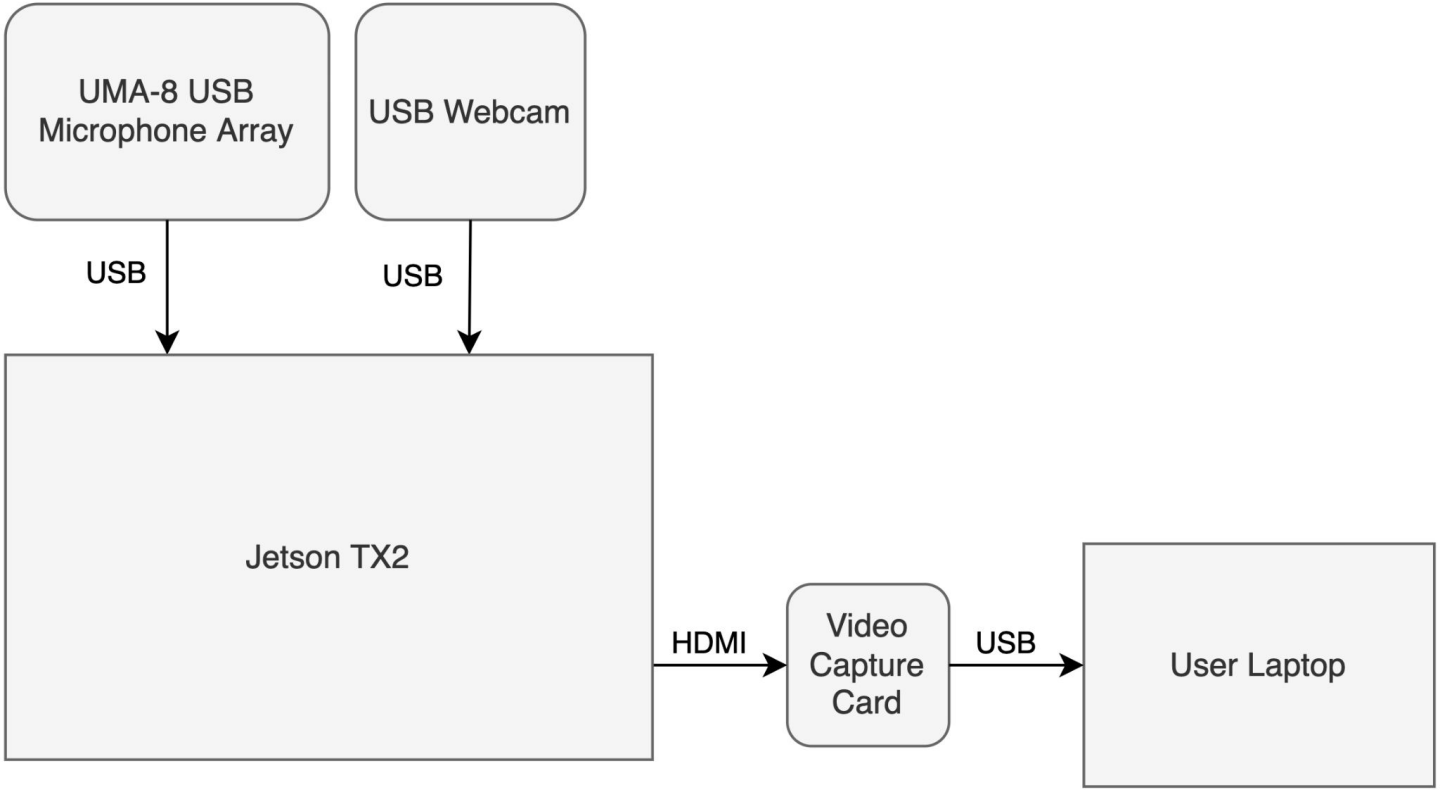
Metric	**Requirement**
Accuracy of speech to text output	Word Error Rate <= 20%, Microsoft standard
Minimum number of speakers	2
Delay: Video+audio capture to captioned video display	< 2s delay, Zoom live captions
Quality of video played for user	720p (HD) 30fps, suitable for live streaming
Range of audio capture by microphones	75 Hz to 4kHz, range of human voice
Size and weight of device	Can fit on conference table, < 4.0lb

Solution Approach

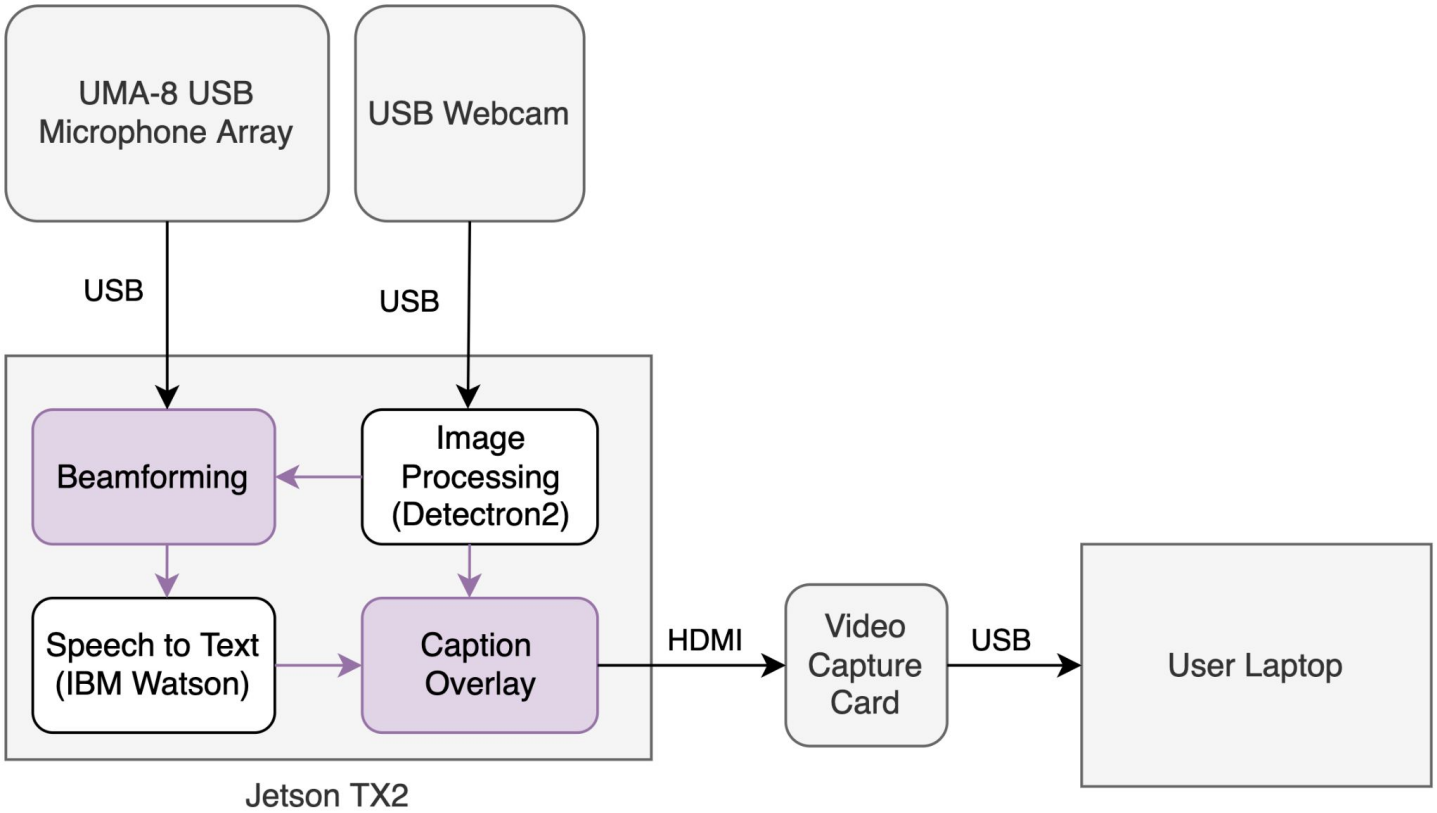


1. Camera identifies location of speakers
2. Microphone array uses beamforming to isolate audio for each speaker
3. Audio is fed to NLP model to produce captions
4. Captions overlaid over final video

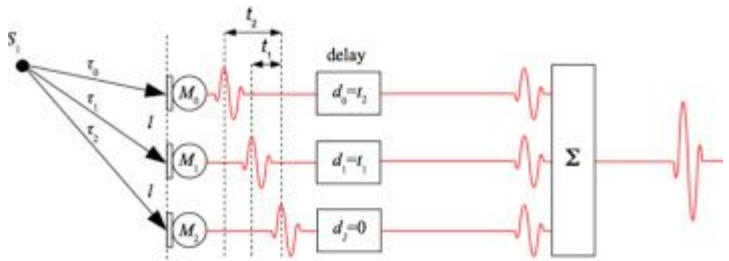
Hardware Specification



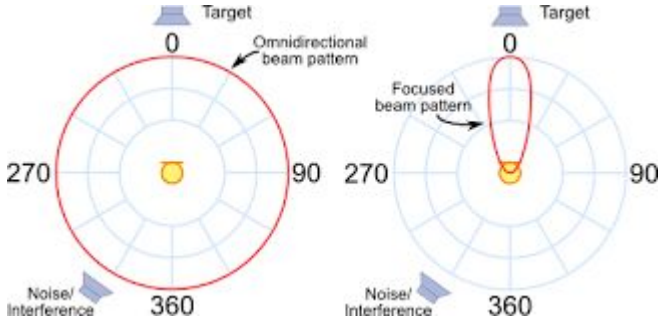
Software Specification

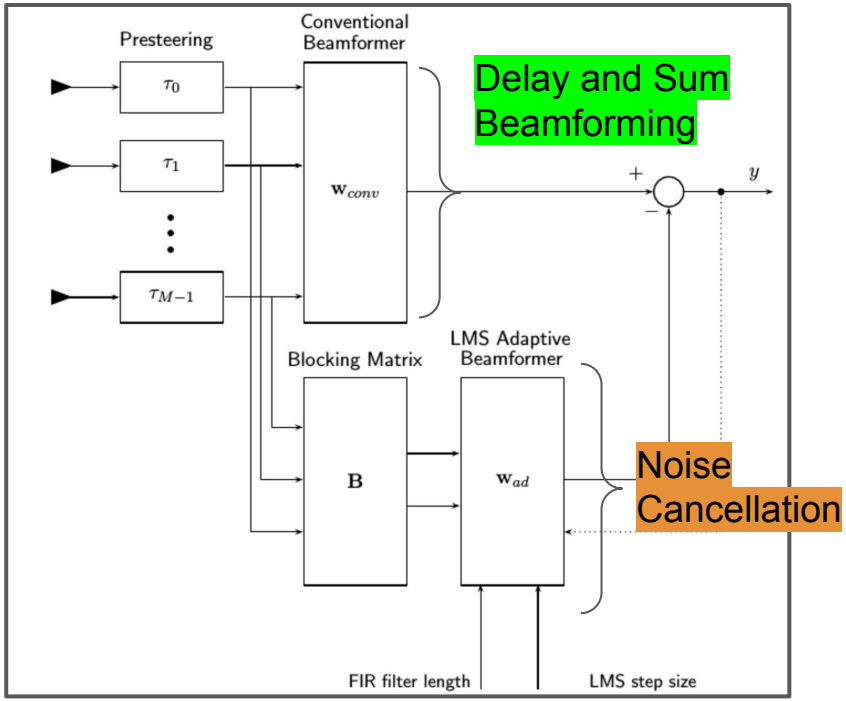


Signal Design



Synchronize the arrival of signals from the same source



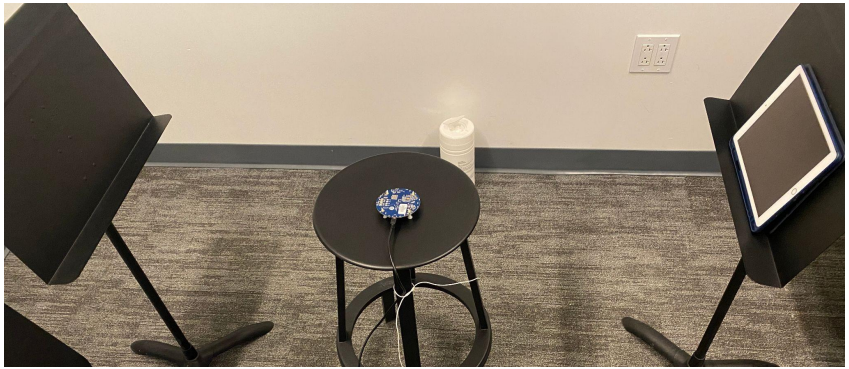


Generalised Sidelobe Canceller

Implementation Plan

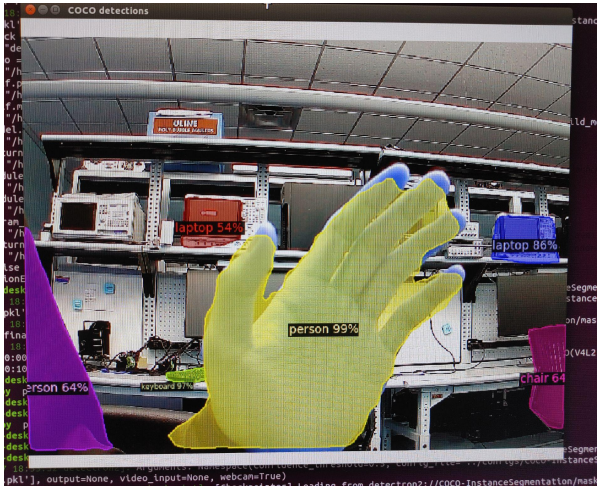
Parts in hand:

- Jetson TX2
- USB Hub
- UMA-8 Microphone array and enclosure
- Wide-angle webcam

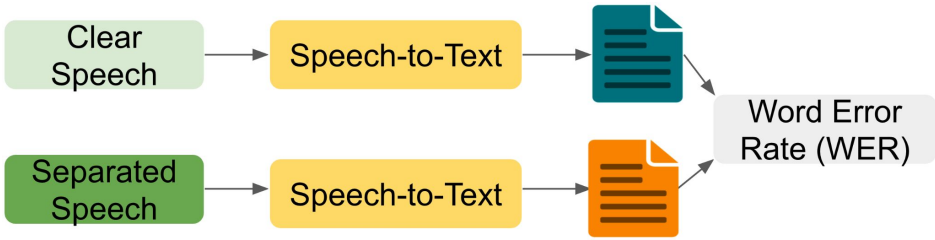


Software tested on TX2:

- Detectron2 image segmentation
- 7-channel microphone recording



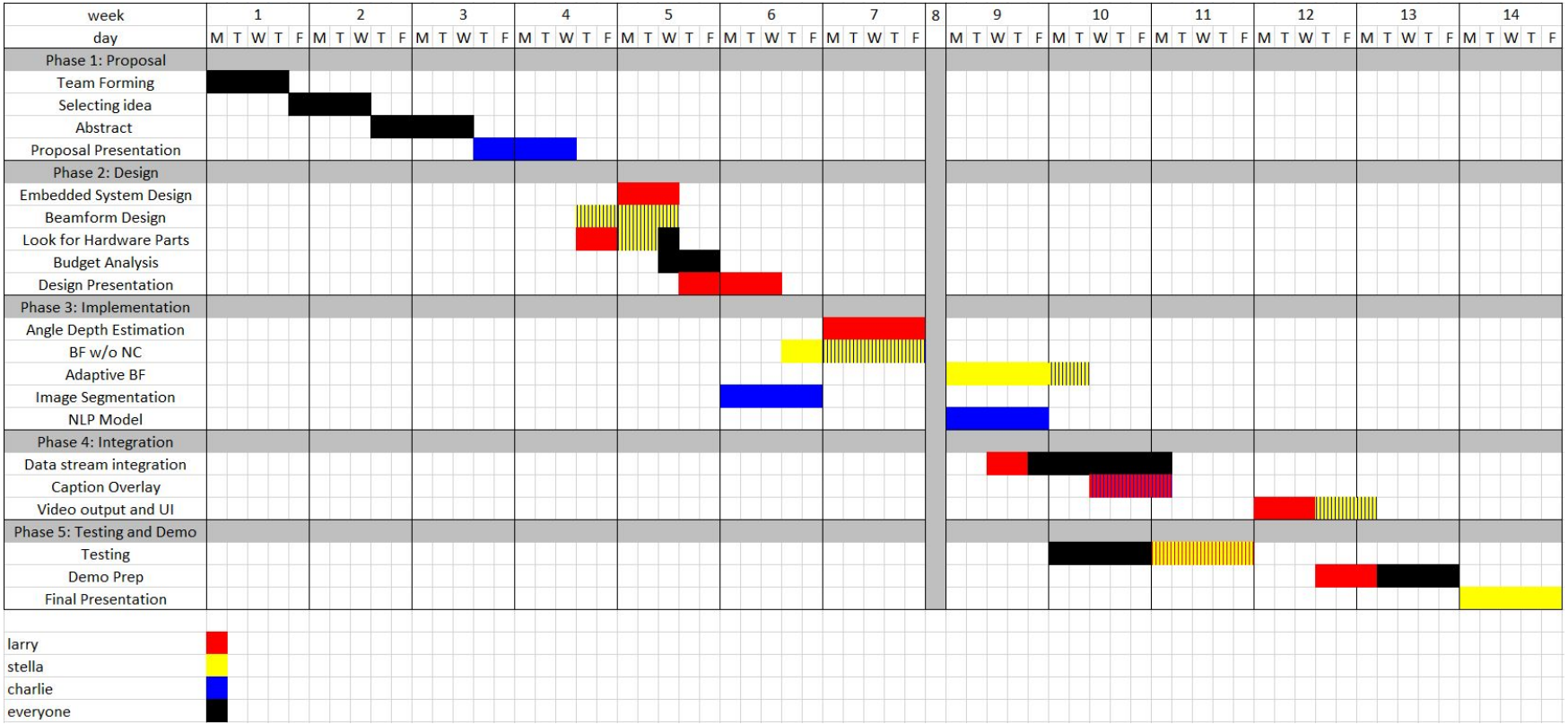
Testing, Verification, and Validation

Metric	Target Value	Test
Accuracy of speech to text output	Word Error Rate <= 20%	
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
Delay: Video+audio capture to captioned video display	< 2s	Record video/audio input and output and measure latency.

Risks and Mitigations

Risk	**Mitigation**
Circular microphone array does not work well with beamforming.	Purchase a linear microphone array or attempt to construct our own.
Beamforming not effective enough for Speech-to-Text model.	Use Sliding Discrete Fourier Transform (real-time STFT) to separate speech based on phase difference.
Processing is not fast enough for real-time applications.	Do processing offline, producing a recorded video as our final product.
Camera estimation of speaker locations is inaccurate.	Integrate angle estimate of sound from microphone array.

Schedule and Work Distribution



Conclusion

We have a specific problem and solution

- Main challenge is beamforming and signal processing
- Simplified surrounding components
- User-friendly and useful output