EyeHear

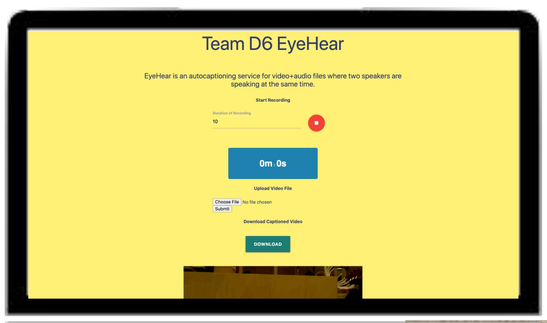
D6: Stella Getz, Larry Geng, Charlie Li
18-500 Capstone Design, Spring 2022
Electrical and Computer Engineering Department
Carnegie Mellon University

Electrical & Computer ENGINEERING
Carnegie Mellon

Product Pitch

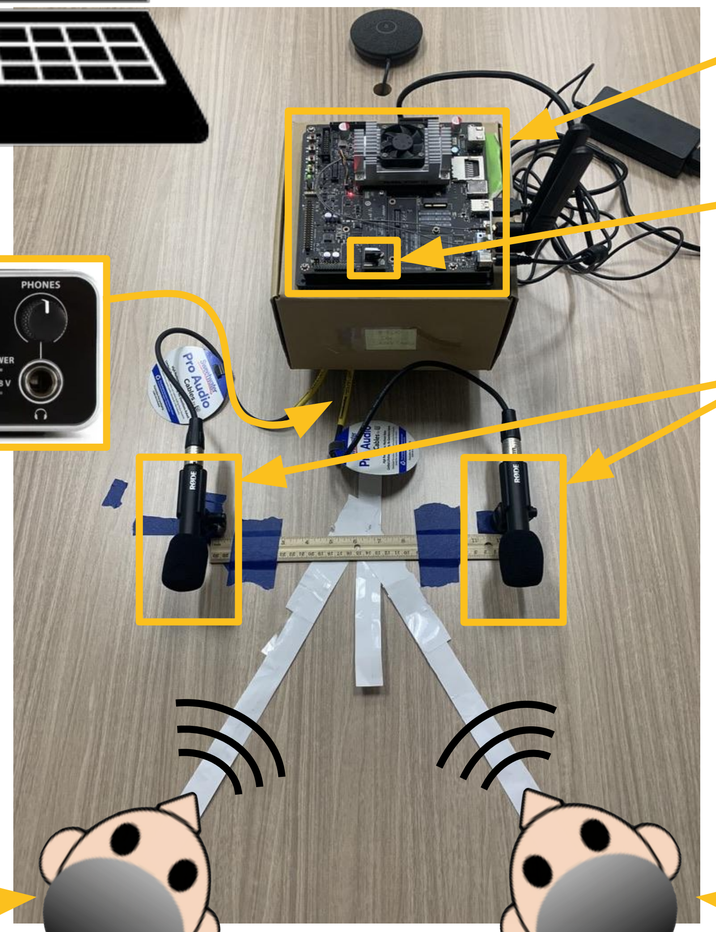
When multiple speakers are speaking at the exact same time, conventional auto-captioning services fail to provide accurate captions of individual speakers. The resulting poor captions can harm viewers' understanding of the conversation. While Google's Looking to Listen speech separation has evolved video based speech separation techniques, it relies on the visual features such as speakers' mouth movement which fails when speakers are wearing masks To tackle this problem, we propose a visual-audio fusion system, EyeHear, aimed at captioning the mixed speech of multiple speakers based on source of sound rather than mouth movements. The system uses off-the-shelves camera and microphone array enhanced with state-of- the-art beamforming, SepFormer speech separation and IBM Watson speech-to-text techniques, to locate and caption the speech of individual speakers.

System Description



User Laptop

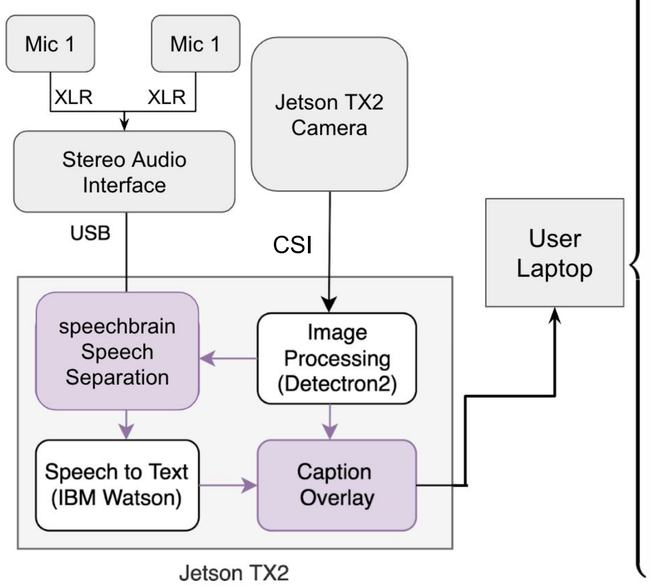
Stereo Audio Interface



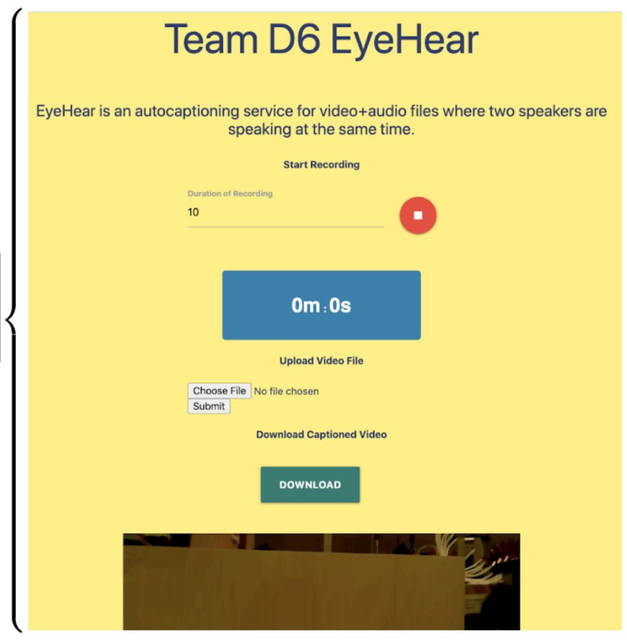
System Architecture

Our system centers around a Jetson TX2 as the main processing component. The TX2 records input from a camera and microphones, and performs four major processing steps: Image segmentation to determine speaker locations, audio separation of the incoming audio, caption generation from the separated audio, and caption overlay onto the input video.

System Block Diagram

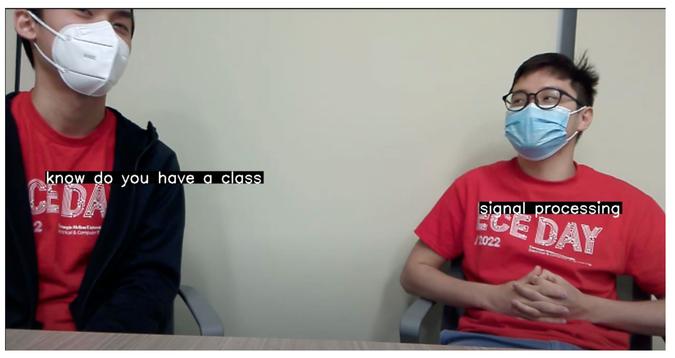


Website:



System Evaluation

We tested our system in a conference room with low noise and reverberation, across various types of speech containing interruptions and overlapping speech. We achieved <=20% WER for most of the scenarios, meeting Microsoft's audible comprehension benchmark.



Frame from a Captioned Conversation

Summary of results across different runs:

Word Error Rate (WER)	Deep Learning Speech Separation										
	take1		take2		take3		average		relative		relative average
Speaker	left	right	left	right	left	right	left	right	left	right	
Left only	6.0	-	10.0	-	4.0	-	6.3	-	-	-	-
Right only	-	22.0	-	18.0	-	24.0	-	22.0	-	-	-
Left starts, Right interrupts	10.0	-	24.0	-	8.0	-	14.0	-	7.7	-	7.1
Right starts, Left interrupts	-	34.0	-	30.0	-	22.0	-	28.6	-	6.6	7.1
Left starts, then Right starts	26.0	34.0	14.0	28.0	40.0	40.0	26.7	34.0	20.4	12.0	20.2
Right starts, then Left starts	58.0	40.0	54.0	26.0	24.0	28.0	45.3	31.3	39.0	9.3	20.2
Left and Right together	20.0	48.0	44.0	30.0	38.0	38.0	34.0	38.7	27.7	16.7	22.2

Conclusions & Additional Information

We developed a robust offline auto-captioning pipeline for different categories of overlapping speech. The novelty of our project is the assignment of separated speech based on deterministic positional information of each speaker, rather than a priori assumption of lip movements. The system we implemented includes most of the features that we initially envisioned. Some missing pieces are real-time captioning and the ability to scale to more than two speakers, which are both possibilities for future work.

http://course.ece.cmu.edu/~ece500/projects/s22-teamd6/

laptop image credit: https://www.shutterstock.com/search/laptop+clipart