# Multimodal Emotion Recognition


Ritwik Das, Luka Eerens, Keerthana PG

Robotics Institute, Carnegie Mellon University


## PROBLEM STATEMENT

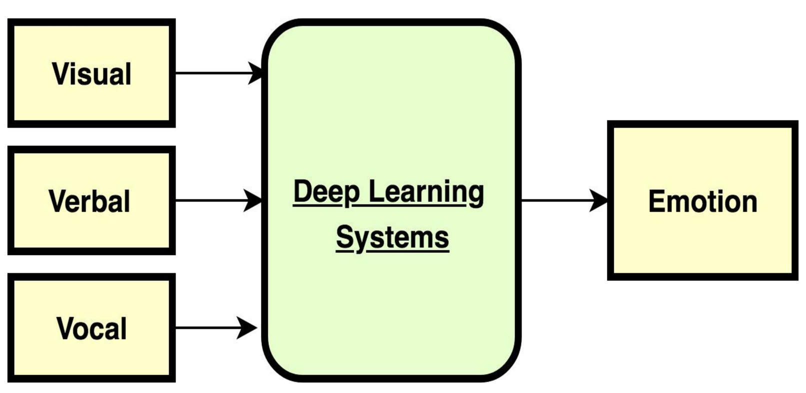


Fig 1: High Level System Overview

Endow a robot with ability to read human emotions: by jointly leveraging visual, verbal, vocal cues.

Visual: Facial Expressions
Verbal: Semantic meaning of words
Vocal: Vocal intonations, loudness, pitch etc.

## USE CASE

Multimodal emotion recognition makes intelligent agents socially perceptive. This helps robots achieve social goals as well as task goals. This technology has multiple applications:

### Home assistants & chatbots

The design of the Lukabot has been guided by chatbots, and home personal assistant bots.

Global home assistant market is projected to grow to $34 billion by 2022 with Amazon selling 15 million Echos in 2.5 years.

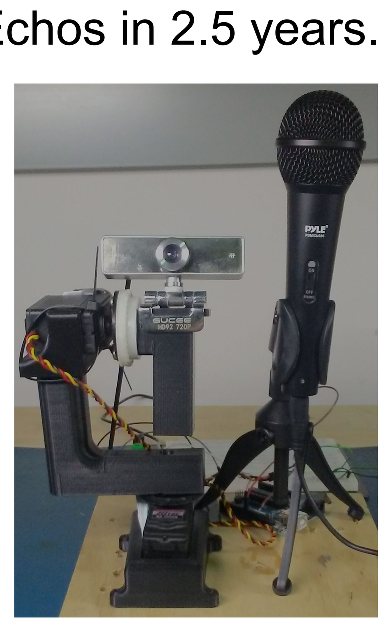

Fig 2: Hardware setup

### Distress detection

Detecting distress can help Lukabot contribute to employee retention, identifying suicidal tendencies and healthcare for the elderly.

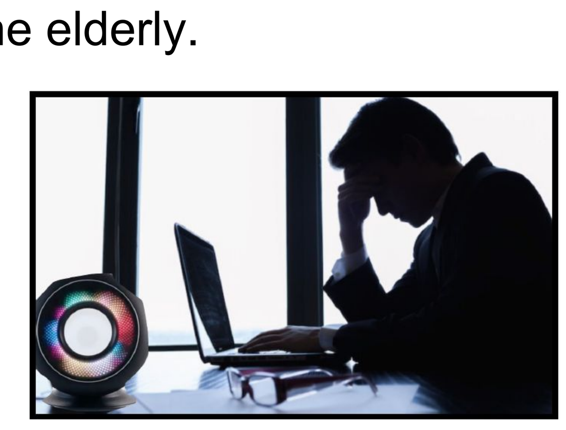

Fig 3: Lukabot in action

## APPROACH & DESIGN

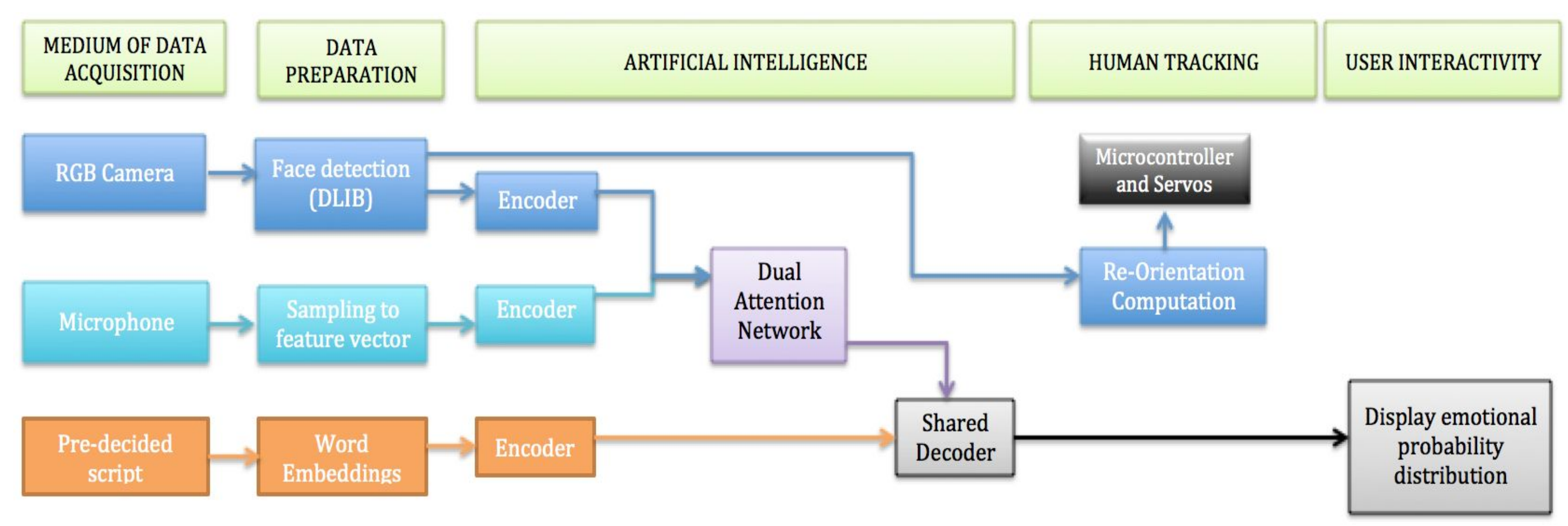


Fig 4: Cyber Physical Architecture

Our system takes in raw images, audio waveforms and a transcript. Pre-processing is done via face detection, wave sampling, and word embedding conversion. These are then passed through encoders and decoders to predict emotion while also tracking the human in real time.

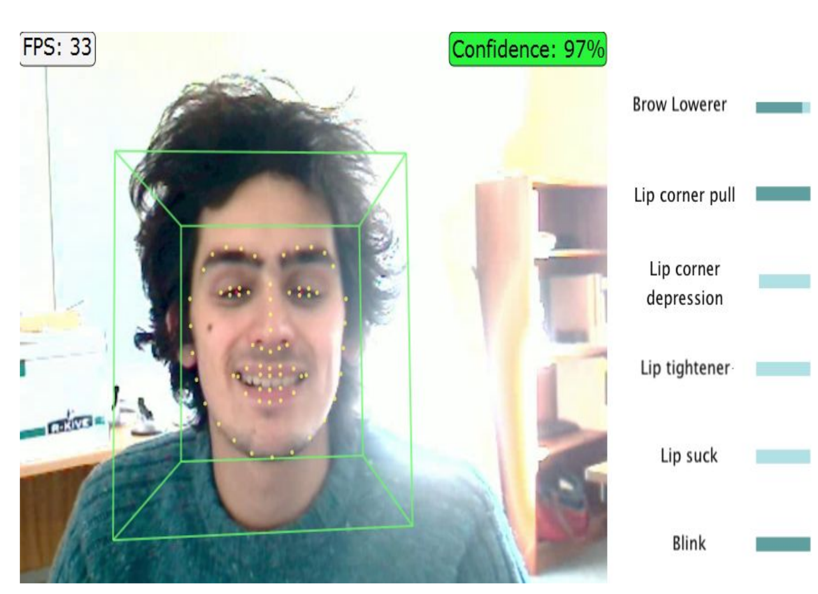


Fig 5: Facial Action Units detection

### Vision Encoder

The vision encoder is used to extract intensity of action units from the face. Action Units(AUs) are fundamental actions of individual muscles or group of muscles. Some examples of AUs include inner brow raiser, lip tightener and dimpler.

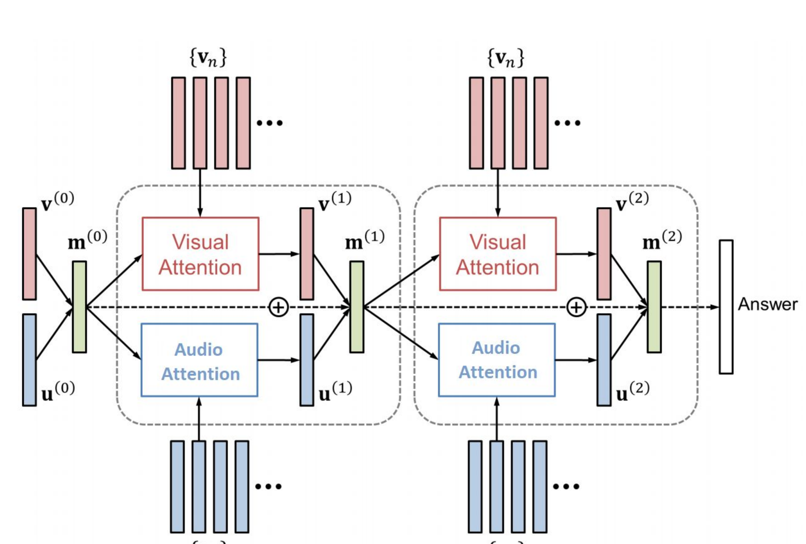


Fig 6: Architecture of Dual Attention

### Dual Attention Network(DAN)

Performs visual and audio attention simultaneously through two sequential steps and gathers necessary information from both modalities. The idea of our DAN is to attend to specific features in both audio and vision modalities. Both the visual and audio attention work by employing soft attention. Our experiments have shown that the Dual Attention is quite robust to noise and corruption of one or both modalities.

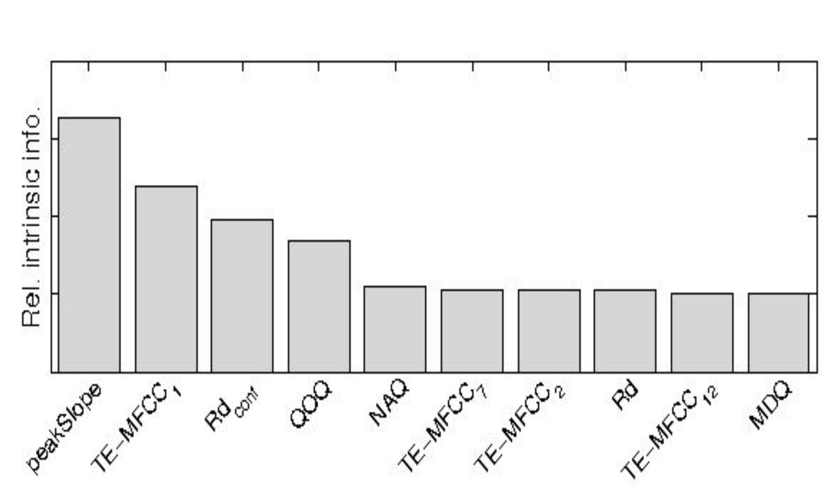


Fig 7: Top 10 audio features

### Text Encoder

The text encoder uses a bi-directional LSTM together with attention to predict emotions. These predictions are then combined with the predictions made by the DAN (above) to give a combined final emotion result.

### Vocal Encoder

Extracts acoustic features like 12 MFCCs, pitch tracking, glottal parameters, peak slope parameters etc. These extracted features carry different characteristics of human voice which have been shown to be related to human emotion.

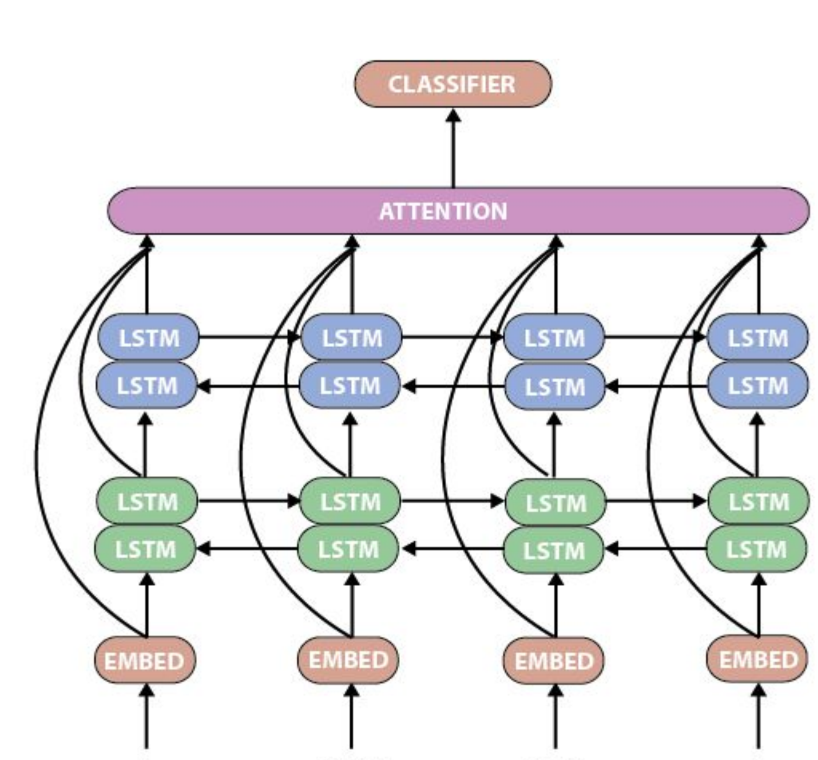


Fig 8: Text Modality Architecture

## RESULTS

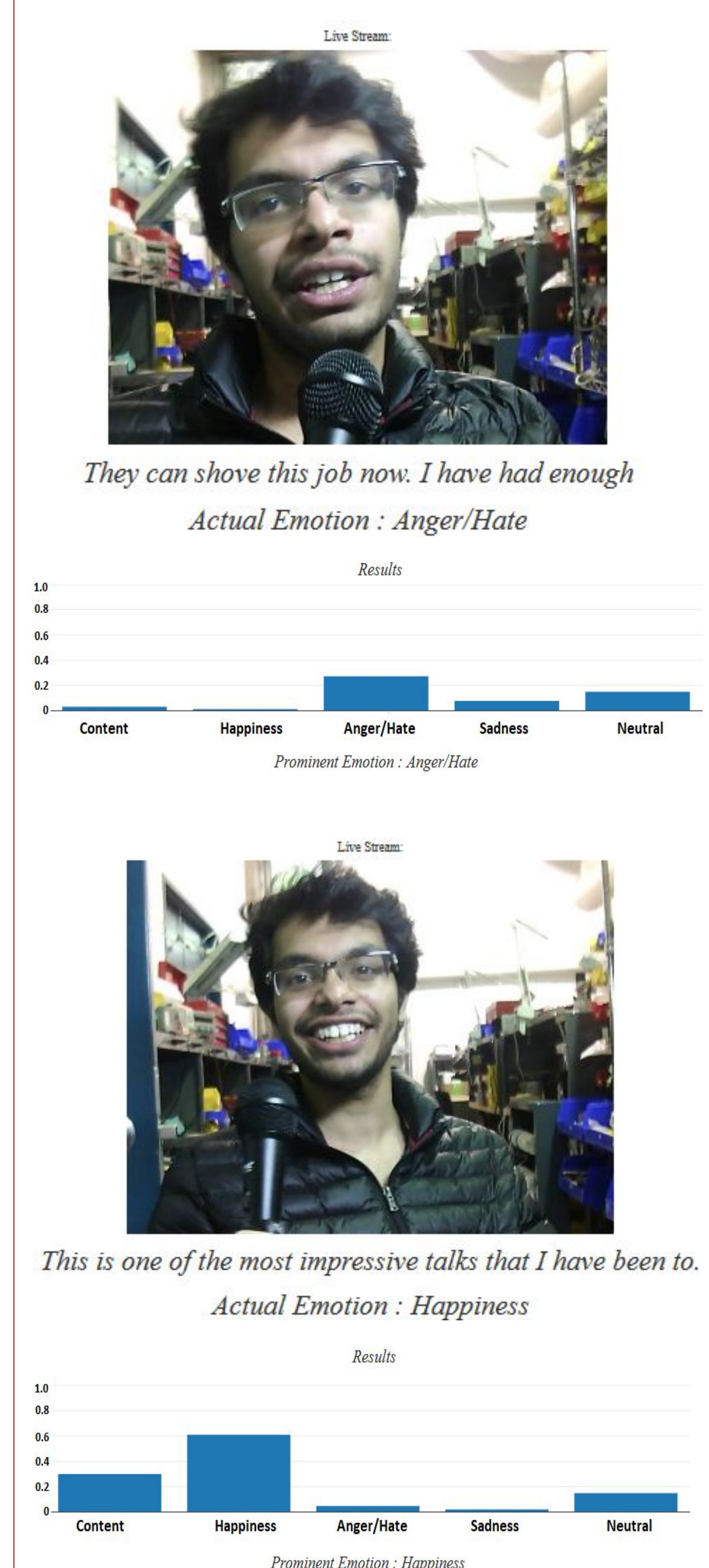


Fig 9: Snapshot of our UI showing predictions

| Modality | Accuracy |
|---|---|
| Audio+Vision | 55 % |
| Text | 59 % |
| Audio +Vision + Text | 71 % |

The table above benchmarks our results on real tests conducted with a webcam, mike and a script. The predictions are compared with human labelings across 5 emotions, these emotions being; content, happiness, anger, sadness and neutral.

Results on CREMA-D dataset (across 6 emotions)

| Architecture | Modality | Accuracy |
|---|---|---|
| Human Performance | Audio | 40.9% |
| Audio Encoder +LSTM Decoder | Audio | 41.5% |
| Vision Encoder +LSTM Decoder | Visual | 54.8% |
| Human Performance | Audio + Visual | 63.6% |
| Both Encoders +Dual Attention | Audio + Visual | 65% |

Results on RAVDESS dataset (across 8 emotions)

| Architecture | Modality | Accuracy |
|---|---|---|
| Audio Encoder + LSTM Decoder | Audio | 41.25% |
| Vision Encoder+LSTM Decoder | Vision | 52.08% |
| Both Encoders + Dual Attention | Audio -Vision | 58.33% |

The Lukabot has also proven to fluidly orient its camera with 2 degrees of freedom towards the face of any human test subject in its field of view that moves around at 20 cm/sec.