Conceptual Design Review
Multimodal Emotion Recognition

Team D: Deeply Emotional

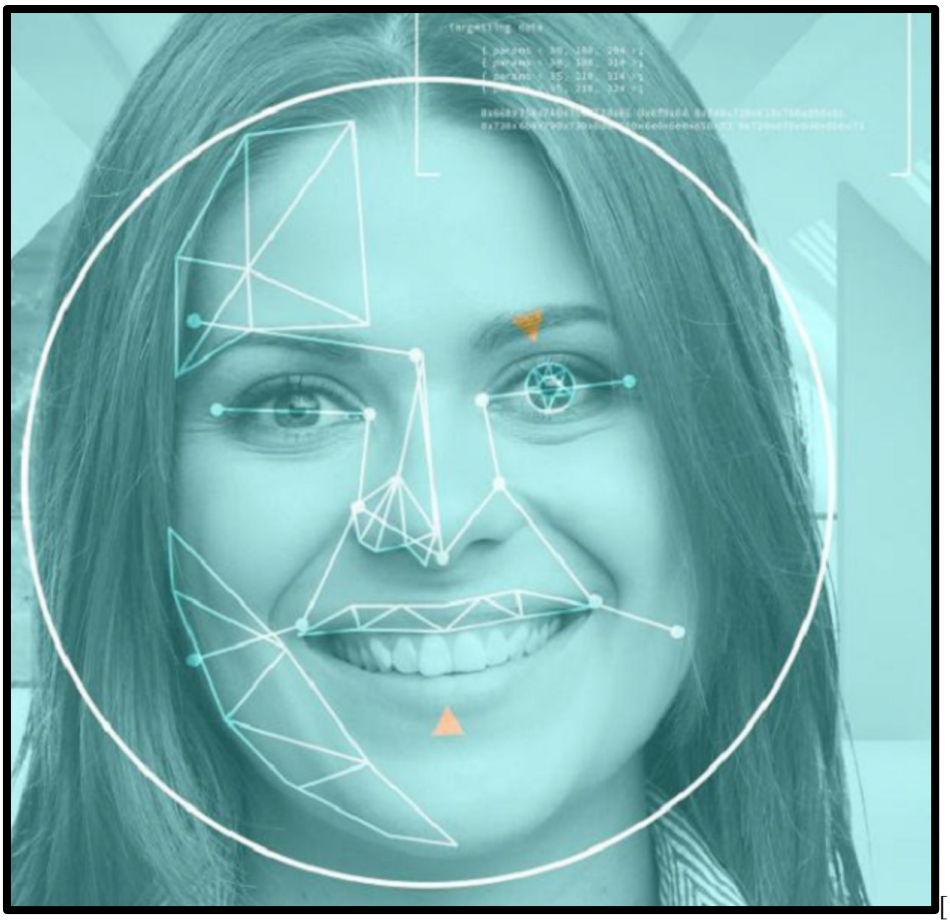
Team Members:

Keerthana P G

Luka Eerens

Luxing Jiang

Ritwik Das



[1]

Table of Contents

1. Prefatory Information

A barrier to the boom of robotics and their integration within the society is their lack of social aptitude. Robotics and AI are seen today as merely tools that assist and not as collaborators and peers because often they ignore the socio-emotional bonds that form the fabric of our society instead of depending on those bonds and personal relationships to improve collaborative skills. AI and robots of the future will likely build inter-personal relationships with the users and for this, they need to be able to detect and respond to emotional cues.

Robots of the future will likely have two goals: a task goal (that accomplishes the robot's designated task) and a social goal (an interaction style is comfortable, sensitive, and results in increased closeness and better working relationship with the user). [2] They would be required to detect, reason and respond accordingly to these social cues.

We are trying to solve a small part of the problem towards this larger goal of social AI. Specifically, we focus on *detecting* social behaviors in conversation, using three modes, verbal, visual and vocal. Our sponsor is Emotech Ltd., a London based startup that aims to build an emotionally aware personal assistant house robot that over time would adapt its personality to yours. As a result, no two Olly would be the same because no two humans would socially interact with their Olly in the same way.

Figure 1: Olly, the world's robot with personality [3]

Key to the success of the vision of Emotech is the ability for humans and their Olly to build rapport with one another in a way that is currently only achievable amongst humans with each other. As such, these Olly will need to not only understand what is said to them but plan their response avenues by factoring in the emotional dynamics of the conversation at run-time.

A nature of human emotion is that it is externalised in a way that it is noticeable by an outside observer. These emotional cues can be observed through, body language, through

facial expressions, through the choice of words in sentences and through the way in which those words and sentences are spoken. Clearly this is a multimodal medium of communication that we as humans are quite good at reading because of our experience with each other over all these years.

The question now becomes: can this ability to read emotions by analysing all these modalities be learnt by machines? And can this be done in an efficient manner that could be deployed into commercial products like Olly?

In order to explore this challenging question we have assembled a team of four enthusiastic MRSD students with deep learning experience to prototype a multimodal emotion recognition agent. We also reached out to leading experts in multimodal machine learning and human psychopathology such as Dr Jeff Cohn and Dr Louis Phillip Morency for advice.

2. Project Description

2.1 Project Details

This project fits under the HMI (Human Machine Interaction) section of MRSD projects. It involves applying machine learning techniques in order to build an AI that can detect human emotions through a multitude of data modalities.

What this project is NOT is an attempt to recognise the mode of incoming data and detect emotions from that single data mode. On the contrary, all data modalities will be fed to the system, just as we humans receive it (through vision, hearing) and the system will need to detect emotions by jointly considering all data modalities.

Another thing this project is NOT is an attempt to build an intelligent chat-bot, nor is it a speech technology project that aims to convert audio data into text. Though these are pertinent to Olly and other emotionally aware personal assistants, they are not related to the title of our project and so their implementation (if needed) will be done with readily available APIs.

We are more focused on the multi-modal aspect of emotion recognition where video data, acoustic waveform analysis, and lexical sentiment analysis is jointly used to predict emotions of the humans in front of a camera and within "earshot" of a microphone.

2.2 Project Goals

The goal of this project is to build a high EQ (Emotional Quotient) AI agent that jointly uses acoustic, lexical and visual information to predict human emotions.

More specifically, this information will be what we humans use to gauge the emotional state of other humans:

- Visual: Facial expression, pose and orientations(smiles, frowns, eye gaze, head nod)
- Vocal: Vocal expressions (Laughter, Groan), Prosody (Tones, Pace, Pitch)
- Verbal: Natural Language and Semantic Sentiment

We thus aim to prototype and test multimodal deep learning systems that sample this Three-V (Visual, Vocal, Verbal) data, and output emotions as close to real time as possible.

The applications of this project are several:

1. Improving perception for and of the socially challenged- This use case is explained in detail below.
2. Enhancing social skills of current robot assistants(eg home robots, virtual assistants, etc)
3. Perceptive and targeted marketing: The tool could be used to gather the response and likability of people towards advertisements and products, and this data can be used for better marketing.

3. Use Case

Michael is a software engineer for a very demanding company and also happens to be extremely shy, has approach anxiety and has little to no verbal exchanges with others at his office. Despite these mind blocks he has in regards to speaking to others, he is introspective enough be aware that he has a problem and commits himself to solving it. In order to practice small talk, Michael instinctively decides to buy the Amazon dot as in his eyes, a robot will not ostracize or judge a person as socially inept as him. A short while later, the dot is delivered and after setting it up, he begins his dialogue. However despite his issues, Michael quickly becomes aware of just how sterile the conversations with the dot are. He is interviewing the dot, which is returning bland, lifeless answers.

Discontent with this purchase he searches the market for alternatives and finds Olly the personal assistant from Emotech. After the order and delivery Michael takes a deep breath and flips the switch. An AI agent comes to life, notices him and orients its robot body towards him and proactively starts a conversation. Michael is shocked, he has already begun to anthropomorphise the robot because of its act of facing him when noticing him, and breaking the ice. He responds and the conversation becomes dynamic. Not only that, the robot seems to choose its words carefully from reading his externalised emotional queues. This is reflected in the proactive suggestions by the robot as well as its responses or lack thereof to Michael's words. The conversation continues, time flies and before he knows it, Michael has had a 30 minute long conversation with the robot where he has vented about his problems, opened up and talked about his life.

Figure 2: Michael with his Olly [4]

These interactions occur every day as Michael gets back home from work. He begins to feel progressively better about himself day by day as this Olly robot provides a vessel to release himself of his psychological troubles by venting to it. Just like a psychologist, Olly listens and guides Michael into appropriate topics from his answers. To an outside observer, these interactions seem to indicate a positive trend in the right direction for Michael. Only a short while ago he was having trouble finding his words during conversations, had little experience conversing with other human beings and was completely incapable of building rapport with anyone. He was also depressed by his interactions with their concomitant missteps, awkwardness and gaffes. Olly seems to have addressed both of those issues: first by being a loyal friend that he can practice having meaningful conversations with, and second by therapeutically letting him vent.

The regular interactions with Olly have allowed Michael to regain his confidence and have honed his ability to hold a conversation. This has led to gradual improvements in the quality of his interactions with his coworkers at the office. He also feels less depressed and this is monitored by Olly as it looks for trends in changes to Michael's overall sentiment in each conversation.

The key driver to the helpfulness of these interactions is Olly's ability to read Michael, and this comes from a strong emotional awareness that was engineering into the robot. Though the opportunities are endless for deep emotion awareness engrained in robots, the use case presented above focuses on assistance to socially lacking humans. This is not a single incidence use case because the benefits only accrue from systematic incidences of conversation occurring over days or weeks.

4. System-level requirements

The following is the breakdown of our project requirements over the two semesters. These requirements are categorized as Mandatory (M) or Desirable (D), as well as Performance(P) and Nonfunctional (N).

4.1 Functional requirements

Vocal data is far less multi-dimensional than image or script, due to this, we feed the vocal data into the network with minimal processing (noise reduction, etc). However, we require more focused feature extraction for

Table 1. Functional requirements details

ID	Title	Description
M.P1	Detect human faces from captured images	The system will detect human faces from images with an accuracy of up to 80%
M.P2	Converts speech to text	The system will convert medium paced legible speech to text with an accuracy of up to 70%
M.P3	Extracts features from script	The system will detect features from script with an accuracy of up to 60%
M.P4	Output emotion chart	The system will output the final emotion recognition result with an accuracy of up to 60% and with a speed of 1 frame/s
M.P5	Track user	The camera will track user in real-time

4.2 Non-functional requirements

Table 2. non-functional requirements details

ID	Title	Description
D.N1	Rests on tabletop	The robot where the system deploys shall rest on tabletop
D.N2	On/off switch	The system shall be standalone and have a switch to activate and deactivate
D.N3	Under $5000	The budget of developing this system shall be under $5000.

5. Functional Architecture

Our functional architecture can be divided into three subsections: visual, acoustic and linguistic.

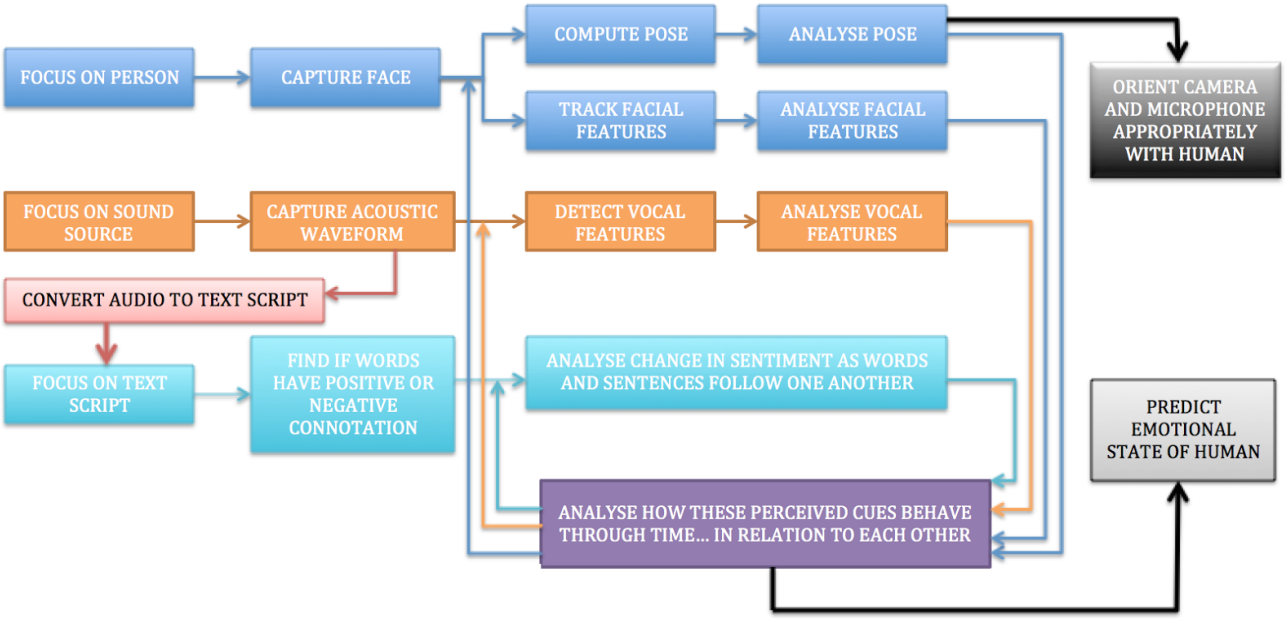


Figure 3: Functional Architecture

The visual section will take in information from human faces. The visual system can also be divided into two parts. In the first part, it can compute and analyse the user's face pose. After computing poses, the camera can tracker users in real-time to make sure that it can capture facial expressions when the user moves. In the second part, the subsystem is to take in facial information and analyse facial features. This subsystem shall output useful features from visual perspectives.

The acoustic section will focus on sound source. This subsystem uses a microphone to capture acoustic waveforms from what users say. The algorithm inside is going to analyse vocal information in acoustic waveforms. This subsystem can output useful audio features to predict human emotion.

The linguistic section will focus on analysing sentiment meaning of user's words. First, this subsystem uses an API(Sphinx) to convert audio information to script. Then natural language processing algorithm is implemented word by word in script to analyse user's sentiment information.

Finally, the multimodal machine learning part is to integrate the output messages from all three modalities. It is also responsible for analysing the temporal relationship between different modalities through time. The ultimate goal of this section is to give a real-time display of human emotion state.

6. System Level Trade Study [5],[6]

We have done trade studies on the following topics in multimodal machine learning:

1. Joint Representation versus Coordinated Representation
2. Multimodal Fusion
3. Datasets
4. Co-Learning

6.1 Joint Representation versus Coordinated Representation:

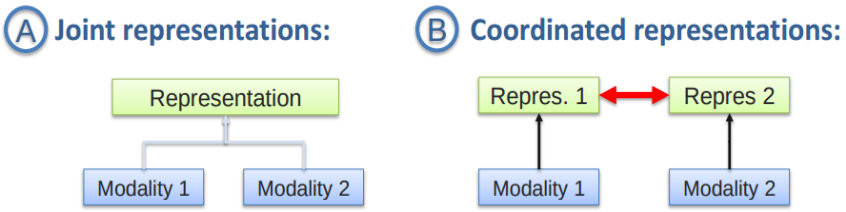


Figure 4: Representation [5]

Joint representations are projected to the same space using all the modalities as input. Coordinated representations on the other hand exist in their own space but are connected to each other through a similarity like Euclidean distance or structure constraint.

Joint Representation:

- Project multimodal data into a single space
- Best suited for situations when all of the modalities are present during inference
- Models can be trained end to end learning both to represent data and perform a particular task
- Cannot handle missing data easily although some ways exist to solve this issue

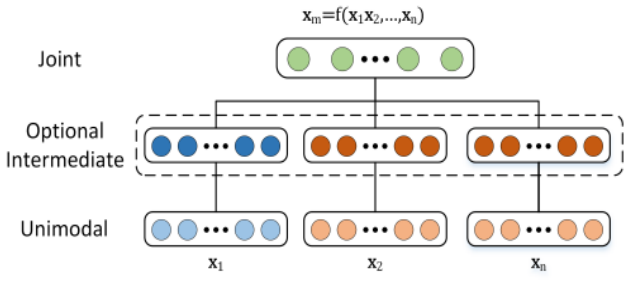
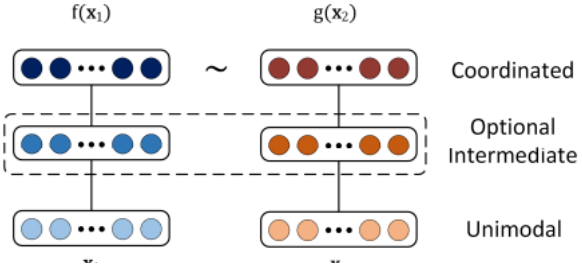


Figure 5: Joint Representation [5]

Coordinated Representation:

- Project each modality into a separate but coordinated space
- Also suitable for applications if only one modality is present during test time
- Such representations have not been worked out for greater than two spaces yet



Conclusion: The reasons for choosing joint representation are owed to

- Existence of more than two modes (Visual, Verbal & Vocal),
- Existence of fully labelled multimodal data set
- Expected multimodal data at real test time scenarios.

6.2 Multimodal Fusion:

Multimodal fusion refers to the joining of information from two or more modalities to perform a classification of classes. Model agnostic approaches of fusion methods are broadly divided into two categories.

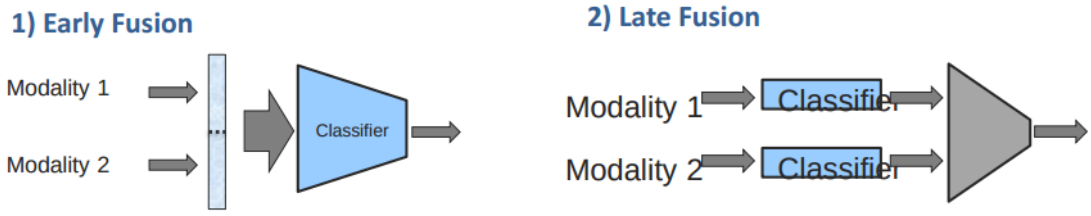


Figure 7: Multimodal Fusion [5]

Early Fusion :-

- Aggregates the features immediately after they are extracted.
- Either simply concatenate the vector representations of the modalities themselves or use an encoder to do so.
- Learns to exploit low level features of each modality
- Faster training pipeline due to the need of training only one network

Late Fusion :-

- Uses single mode classifiers
- Fusion at the end using a weighted average or learned distribution scheme.
- Ignores low level interaction between the modalities.
- Slower training pipeline due to 3X networks and 3X weight matrices.
- Can learn or predict when one or more modes are missing.

Conclusion : - Early fusion was chosen over late fusion because of:

- Faster training pipeline(due to project time line constraints).
- Faster iteration and reiteration due to the above.
- Availability of all three modes both during training and testing.
- Ability to exploit low level features of each mode.

6.3 Data-set

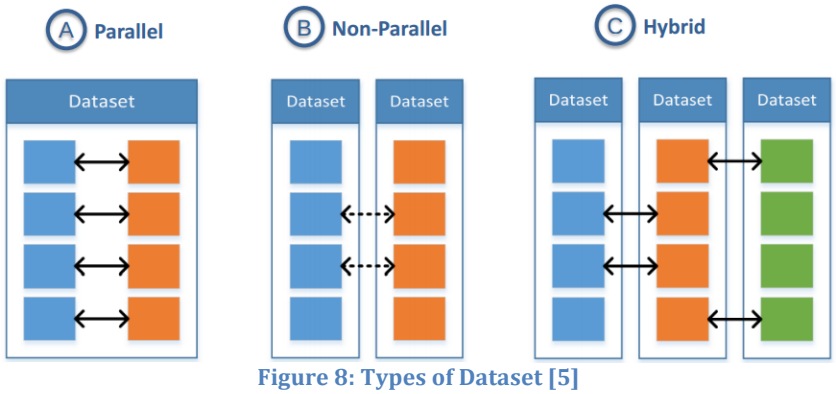


Figure 8: Types of Dataset [5]

Parallel Data: All modalities are in a singular dataset. They have a one to one correspondence.

Non-Parallel Data: Unimodal instances are independent of each other. Typically used for learning representations.

Hybrid Data: Two non-parallel modalities are bridged by a shared modality.

Conclusion: We chose the parallel data-set owing to the following reasons:

- Since our main goal is to predict emotion, concerning with interpretability of data is unnecessary.
- Availability of parallel data-set(s).
- Less complexity and needs only training a single network.

We have also conducted a dataset trade study by reviewing the landscape of existing pertinent datasets from the following source [7]:

Table 3: List of Datasets we have reviewed

Identifier	Modalities	Emotion elicitation methods	Size	Nature of material
HUMAINE	Audiovisual+ Gesture	Naturalistic and Induced Material	50 Clips ranging from 5 seconds to 3 minutes. labelled for emotional content	Mainly interactive discourse
Belfast Naturalistic Database (Douglas-Cowie et al 2000, 2003)	Audio- visual	Natural: 10-60 sec long 'clips'	125 subjects; 31 male, 94 female	Interactive unscripted discourse
Chung (Chung 2000)	Audio-visual	Natural: television interviews in which speakers talk about sad and joyful moments in their lives	77 subjects; 6 1 Korean speakers, 6 Americans	Interactive unscripted discourse
SALAS database	Audio-visual	Induced: subjects talk to artificial listener & interact with different personalities of the listener	Pilot study of 20 subjects	Interactive discourse Subjects unscripted Machine scripted
ORESTEIA database (McMahon et al. 2003)	Audio + physiological (some visual data too)	Induced: subjects encounter various problems while driving	29 subjects, 90min sessions per subject	Non interactive speech: giving directions, giving answers to mental arithmetic etc
Belfast Boredom database (Cowie et al. 2003)	Audio-visual	Induced	12 subjects: 30 minutes each	Non interactive speech: naming objects on computer screen
XM2VTSDB	Audio-visual	N/A	295 subjects ; Video	High quality colour images, 32 KHz 16-bit sound files, video sequences and a 3d Model, a total of 2,360 images

Polzin (Polzin, 2000)	Audio	Acted: sentence length segments taken from acted movies	Unspecified no of speakers. Segment numbers 1586 angry, 1076 sad, 2991 neutral	Scripted

6.4 Co-Learning

Co-learning is aiding the modeling of a (resource poor) modality by exploiting knowledge from another (resource rich) modality. It is particularly relevant when one of the modalities has limited resources — lack of annotated data, noisy input, and unreliable labels. Since we have a parallel data-set, there are two main class of algorithms to achieve co-learning:

Co-training :

- Creating more data from a limited labeled training data. Semi-supervised method of co-learning.
- In an intuitive way it is another name for data augmentation in multimodal machine learning.
- It may cause over-fitting due to over-sampling of same data.

Transfer learning(in the context of multimodal machine learning)

- Mapping from one modality into another modality.
- Typically using an auto-encoder or a Deep Boltzmann machine leads to better multimodal representation functions.
- Can discard noisy modes and use more information from other modes during real - test time.

Conclusion:

Transfer learning was chosen as the method for co-learning due to the following reasons:

- Architecture must handle noise during real test time in Spring Validation Experiment.
- Currently, the team doesn't see a lack of labelled data. If in the future the bias error is on the higher side, this trade-study needs to be revisited.
- Availability of unlabelled data with Prof. Jeff Kohn has also put this trade study into further consideration later on if desired.

7. Cyber-physical architecture

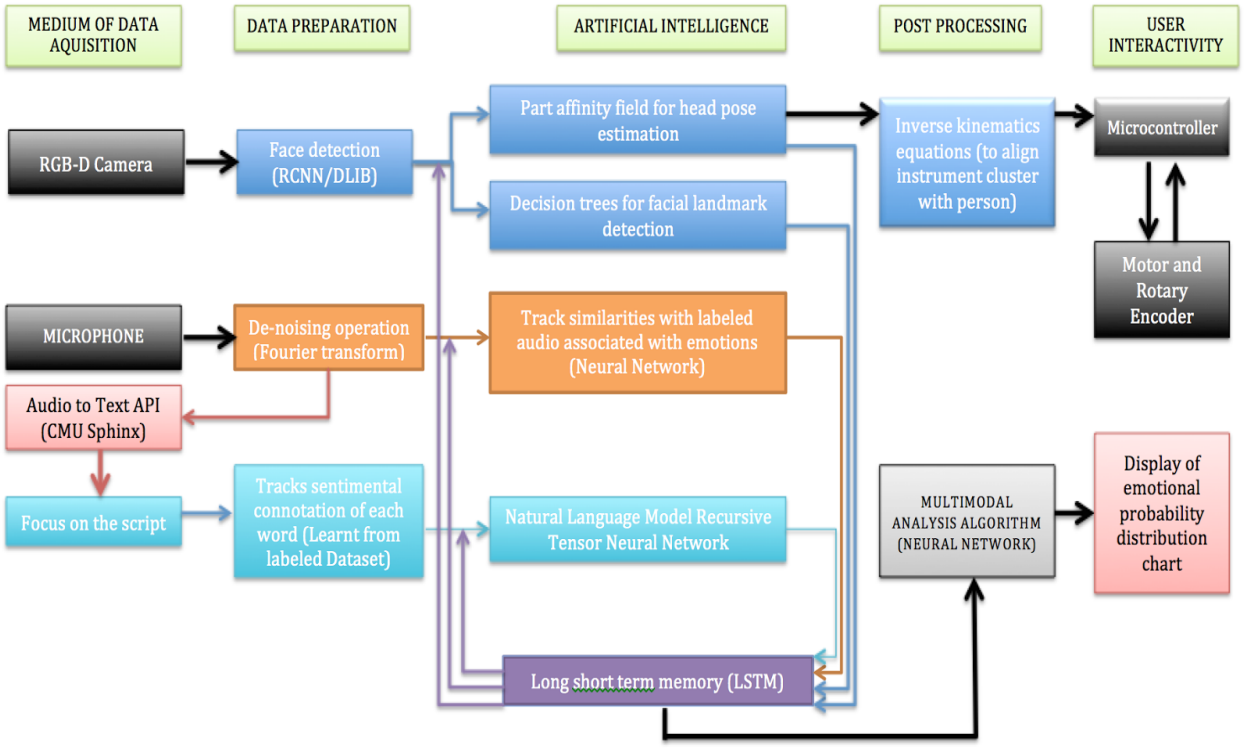


Figure 9: Cyberphysical Architecture

The cyber-physical architecture of project's intended solution is presented above

Since Emotech will provide a home robot called olly, this project mainly focuses on software and has a limited hardware with some sensors. This cyber-physical architecture can also be divided into three parts: visual part, acoustic part and linguistic part. And each part goes through the same five phases: data acquisition, data preparation, artificial intelligence, post processing and use interactivity.

For the visual part, the sensors used to capture images is a RGB-D camera attached to the robot. Since the visual subsystem focuses only on human faces, the dlib library is used to detect human faces in these captured images. This pose subsystem use part affinity fields method to do real-time head pose estimation. This is helpful for the microcontroller to let camera rotate in two dimensions and do real-time tracking of users. For the facial expression part, the subsystem use decision trees for facial landmark detections to get useful features for recognizing human emotions from the visual perspective.

For the acoustic section, a microphone is installed to capture acoustic waveforms of human's words. The first step to process acoustic information is to implement fourier transform to remove noise. Then the subsystem uses recurrent neural network to analyse the similarities between labelled audio and emotion. Finally the neural network can output useful acoustic features for emotion recognition.

For the linguistic section, the denoised acoustic information in the previous part is utilized to convert to script. The conversion process can be achieved with high accuracy by an API like CMU Sphinx. Then a natural language model built by Recursive Tensor Neural

Network is used to analyze sentiment information from converted script. The output of this neural network should be the sentiment state of the user.

Finally, the multimodal machine learning part use neural network to analyze temporal relationships between different modalities through time and integrate the output messages from all three modalities. The final emotion probability distribution output is displayed lively in a screen.

8. Subsystem Description

8.1 Visual Recognition Subsystem

8.1.1 Camera and Motors

RGB-D camera is used to capture sequences of images, the images information can be transmitted to the computing unit for further processing. In order to track users in real-time, a microcontroller and motor are utilized to rotate the camera in two dimensions(up and down, left and right), according to the pose estimation results.

8.1.2 Pose estimation system

The visual subsystem focuses particularly on face features. So the dlib library is used to detect human faces in these captured images. There are two ways as to how the pose estimation system is planned to be designed. One way of doing this is to use a CNN to predict the pose of the face at every frame. Another way is to use part affinity fields(PAFs) to output head pose estimation in real-time. A PAF is a 2D vector field for each part. For each pixel in an area belonging to a particular region the 2D vector encodes the direction that points from one part to the other. This output result is useful both in aligning camera with human faces and emotion recognition.

Figure 10: Results of dlib landmark detector [8]

8.1.3 Facial estimation system

Facial landmark is very useful for emotion recognition. The most important part of this subsystem is using machine learning algorithms to get useful features from facial landmarks detection. These visual features are valuable input for multimodal machine learning. Through a classifier, it can be be a standalone subsystem to recognize human emotions from visual perspective.

8.2 Acoustic recognition subsystem

8.2.1 Microphone

The acoustic information from users is captured by a microphone installed in Olly and transmitted to computing units for further processing. Usually, there is much noise in acoustic waveforms. So the first step to process acoustic information is denoising. The denoised information can go through further processing.

8.2.2 Audio processing system

The goal of this subsystem is to output useful acoustic features for further emotion recognition.

8.3 Linguistic recognition subsystem

8.3.1 Audio to text transcription

This part will use a open source API to translate from audio information to script. This translation will be real-time and robust. The team is currently looking into Sphinx API developed by CMU as their audio to script converter.

8.3.2 sentiment analysis system

A natural language model is built (possibly by Recursive Tensor Neural Network) to perform sentiment analysis on converted script. This model should analyze individual words and phrases in text with respect to different emotional scales. The sentiment output is a useful resource for multimodal machine learning input.

8.4 Multimodal deep network

This part forms joint representations from all three modalities. A single deep network is implemented to give a robust emotion recognition result. There are some important requirements in this part. The first one is this algorithm should be able to analyze temporal relationships between different modalities through time. Multiply stacked LSTMs are one approach to solve this problem. Also, the algorithm should solve the conflict problems between three modalities to give a more reliable prediction.

8.5 Display screen

The display screen shall output six emotions as a probability distribution as well as the dominant emotion. The six emotions are happy, sad, anger, fear, disgust and surprise. The GUI displays human emotion state in real-time. An on/off switch is used to activate and deactivate the whole recognition system.

9. Project Management

9.1. Work Plan

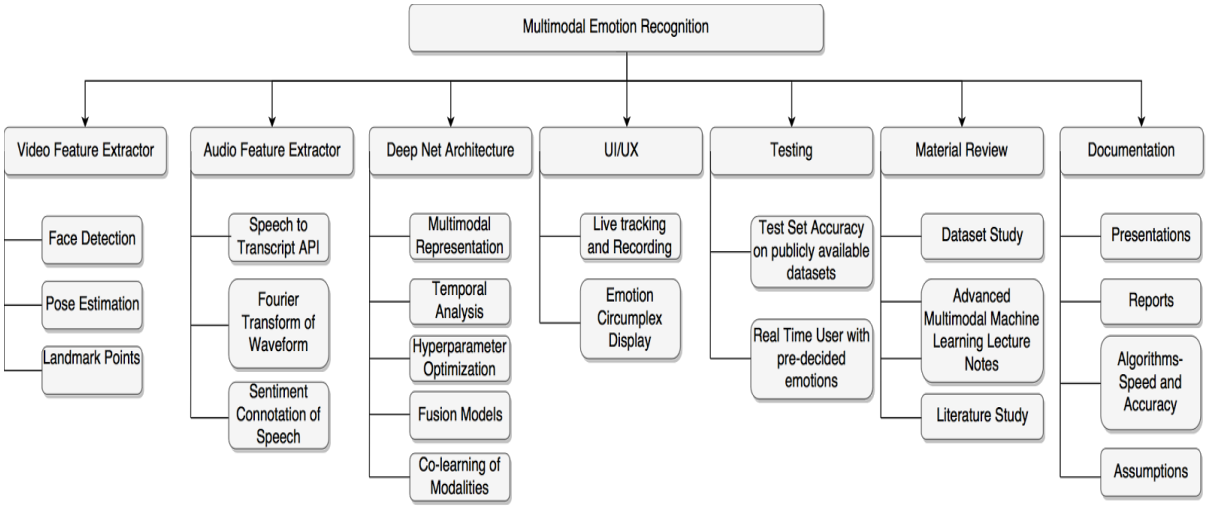


Figure 11: Work Plan

9.2. Schedule

Task Name	30/09/2017	15/10/2017	30/10/2017	14/11/2017	29/11/2017	14/12/2017	29/12/2017	13/01/2018	28/01/2018	12/02/2018	27/02/2018	14/03/2018	29/03/2018	13/04/2018	28/04/2018
Audio Feature Extraction															
Speech to Text API						FVE									SVE
Acoustic Denoising						FVE									SVE
Sentiment Connotation of Speech						FVE									SVE
Video Feature Extraction															
Face Detection	COMPLETE					FVE									SVE
Pose Estimation	COMPLETE					FVE									SVE
Landmark Points	COMPLETE					FVE									SVE
Deep Net Architecture															
Multimodal Representation	COMPLETE					FVE									SVE
Temporal Analysis						FVE									SVE
Hyperparamter Optimization						FVE									SVE
Fusion Models						FVE									SVE
Co-learning of modalities						FVE									SVE
UI/UX															
Live Tracking and Recording						FVE									SVE
Emotion Circumplex Display						FVE									SVE
Testing						FVE									SVE
Test Set on publicly available datasets						FVE									SVE
Real Time User following a script						FVE									SVE
Material Review						FVE									SVE
Dataset Study						FVE									SVE
Literature Review						FVE									SVE
Review of AMML Notes						FVE									SVE

Figure 12: Schedule

9.3. System Validation Experiments

9.3.1 Fall Validation Experiment

Test Conditions

1. Just a test on a test set

Performance Criteria:
Head and face tracking

Procedure: A single user in frame moves in front of the camera
Performance metrics: Accurately track fiduciary points on head and face in different orientations and resolutions

Voice transcription
Procedure: A single user speaks in legible English and reasonable pace
Performance metrics: The speech is transcribed to text with 90% accuracy as compared to average human ability

Emotion Recognition
Procedure: Run on test data and run on real time data
Performance metrics: At least 60% accuracy on test data

Time delay and computing speed
Procedure: Test software on real time data
Performance metrics: Outputs values within a maximum delay of 1 second and computes emotion in 0.9 seconds.

9.3.2 Spring Validation Experiment

Test Conditions

1. Indoor room
2. Lighting conditions from 4000 lux to 5000 lux
3. Single active subject with their own microphones
4. Single person speaks in English from a distance no further than 3 meters

Performance Criteria:

Head and face tracking mechanism
Procedure: A single user in the camera frame moves within the field of view of the camera and a mechanism that rests on a table-top rotates the camera to keep the user centred.
Performance metrics: Ability to maintain the human centered in the field of view while ensuring that their face is clearly visible all of the time.

Voice transcription
Procedure: A single user speaks in legible English and reasonable pace into their microphones
Performance metrics: The speech is transcribed to text with 60% accuracy as compared to average human ability.

Emotion Recognition
Procedure: Run on test data and run 2 different kinds of real time experiments and one test set. The first of the real world tests is a subjective test, the other is an objective test.
Performance metrics: Outputs emotions with at least 60% accuracy for the non real world test set. Outputs emotions with at least 60% accuracy on pre-decided emotions that are scripted by our team for the objective test. Outputs emotions that deviate no further than 20% from subjective human graders for the real world test.

Time delay and computing speed
Procedure: Test software on real time data
Performance metrics: Outputs values within a maximum delay of 1 second and computes emotion in 0.9 seconds.

9.4. Individual Responsibilities

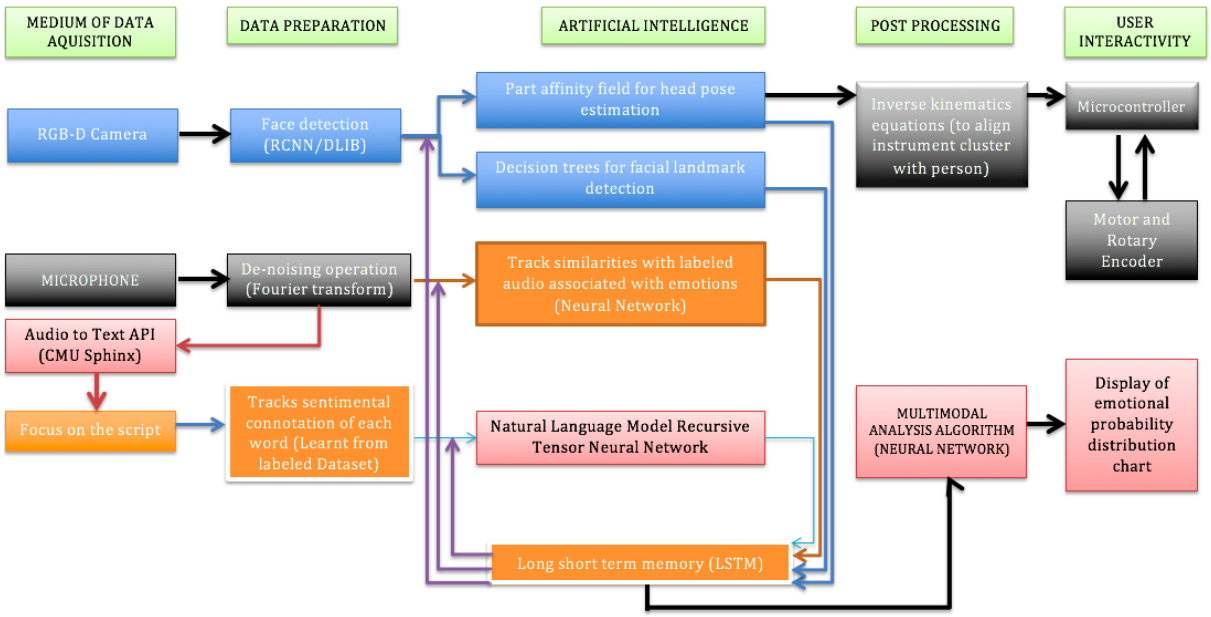


Figure 13: Individual Responsibilities color coded into the cyberphysical architecture

The responsibility of each group member is marked with a particular color as above.
Ritwik: blue, Luxing: orange, Luka: black, Keerthana: pink

9.5. Provisional parts list

Table 4: Provisional Parts List

Part	Purpose	Quantity	Unit Cost	Total Cost
RGB-D Camera	Stereo Cam	1	$100	$100
Microphone	Microphones to capture sound	3	$30	$90
NVIDIA GTX 1080	GPU Computing	1	$500	$500
Arduino Uno	Microcontroller	1	$32	$32
Battery	Battery for motors	3	$30	$90
Motors	Motors	3	$30	$90
Screws, nuts, wires, standoffs	Mechanical design	-	$200	$200
Materials for links and joints	Mechanical design	-	$200	$200

Extras	Extra parts in case of damage/malfunction	-	$500	$500

9.6. Risk management

9.6.1 Technical Difficulties

Table 5: Technical Difficulties

Sl. No.	Description	Likelihood of Occurrence	Level of Impact	Area of impact	Handling Strategies
1.	Lack of large annotated datasets	Medium	High	Accuracy	Survey the landscape of datasets, require permissions
2.	Lack of varied(age, gender, race) subjects in multimodal datasets	High	Medium	Adaptability, Performance	Familiarize networks on varied subjects for single modalities and adapt to multi-modal scenario
4.	Time stamping and synchronising	Medium	High	Performance, Accuracy, Time	Annotate time stamps for each modality, process at the speed of the modality with highest processing time
5.	Delayed output due to high processing time	High	Medium	Time	Optimise code, shift to C++ after prototyping in python
6.	Inadvertent chaos and malfunctioning during demo	High	High	Performance, accuracy, time, grades	Prepare early, use version control, rehearse for demos

9.6.2 Scheduling Difficulties

Table 6: Scheduling Difficulties

Survey. No.	Description	Likelihood of Occurrence	Level of Impact	Area of impact	Handling Strategies
1.	Delay in acquiring permissions and permission denials	High	High	Milestones, scheduling	Look for datasets with minimal permission requirements. Start early and contact a wide range of data set owners.

2.	Sensors and components do not arrive on time	Low	Medium	Schedule	Research readily available components within budget, anticipate requirements and place orders early
3.	Project milestone unlikely to be met	Medium	High	Schedule, performance	See if team can put in more hours or move deadlines or downsize requirements
4.	Part breaks and is out of stock before the demo	Medium	High	Performance	Try to borrow a replacement from CMU or sponsor
5.	Fabrication of parts take too long	Medium	Medium	Schedule	Work in parallelized manner, start early

9.6.3 Personnel Difficulties

Table 7: Personal Difficulties

Survey. No.	Description	Likelihood of Occurrence	Level of Impact	Area of impact	Handling Strategies
1.	Lack of in-depth technical advice	High	High	Competency, Efficiency	Talk to professors at CMU, and research scholars at Emotech
1.	Team is inexperienced in natural language processing	Medium	High	Competency, performance, delays	Consult with outside experts, delegate some tasks
2.	Communicating with remote sponsor	Low	High	Performance, delays	Schedule regular Skype/Hangouts meetings. Meet in person if possible. Give regular updates.
3.	Team member falls ill for long period of time	Low	Medium	Delays	All other team members share tasks
4.	Disagreement among team members	High	Medium	Performance	Use compassion and rational conversation
6.	Miscommunication among the team members	Medium	High	Performance	Have frequent meetings, discuss issues and keep everyone on same page

10.References

[1]https://bismart.com/wp-content/uploads/2016/04/Emotion_recognition03.jpg
[2]http://articulab.hcii.cs.cmu.edu/projects/sara/
[3]https://www.androidheadlines.com/wp-content/uploads/2017/01/Emotech-Olly-official-release-KK-2.jpg
[4]https://g68qpy3g1w-flywheel.netdna-ssl.com/wp-content/uploads/2014/07/depressed-worker-e1405446545412.jpg combined with a screenshot of Olly from this video: https://www.youtube.com/watch?v=xCS8YXhT7j0
[5]**arXiv:1705.06**

[6]Dr. Luis Morency, 2017, Introduction, lecture notes, Advanced Multimodal Machine Learning 11-777, Carnegie Mellon University delivered 29 Aug 2017.
[7] http://emotion-research.net/wiki/Databases
[8]https://matthewearl.github.io/2015/07/28/switching-eds-with-python/