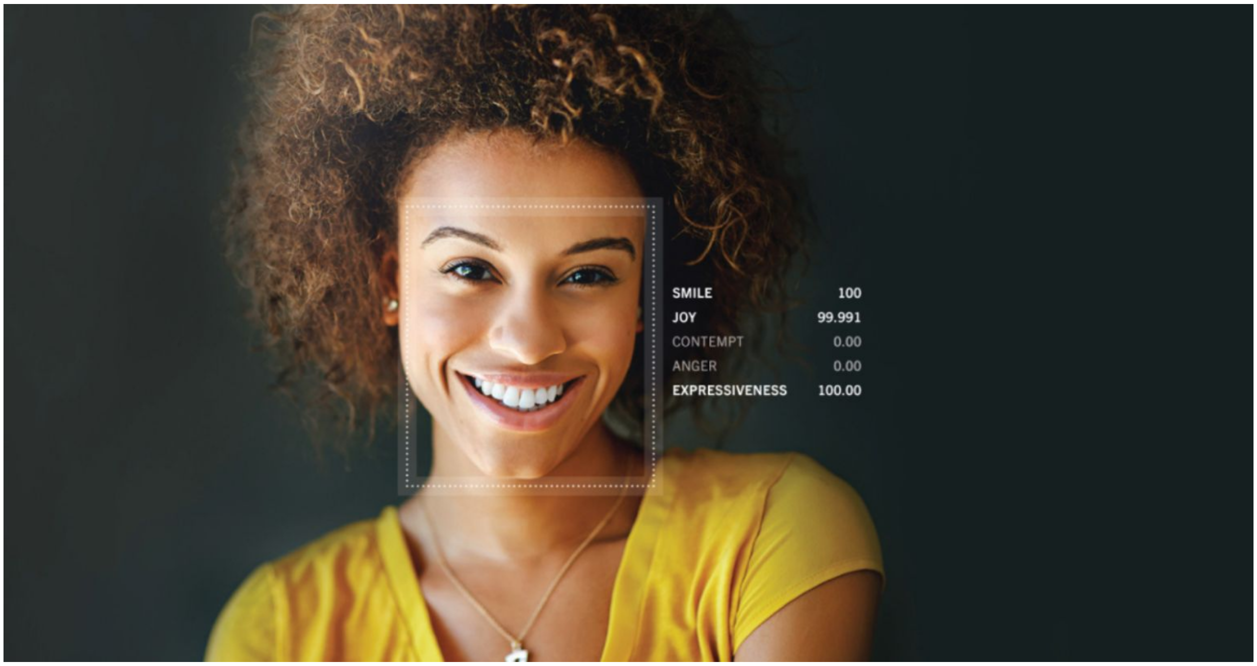
## 16681 - MRSD Project Course

# Critical Design Review Report

14th December 2017

## *Team D: Deeply Emotional*

**Members:**
Luka Eerens
Keerthana P G
Ritwik Das
**Sponsor:**
Emotech Ltd



# Abstract

This projects seeks to build a robotic system that can detect emotions from humans at real time using verbal, visual and vocal modes of data. At our project deep learning algorithms meet robust face tracking hardware to get the best performance. This critical design review jots down the refined overview of our task: highlighting our progress, documenting the status and state of system and cyber physical architectures and tracking various aspects of project management. Additionally, we enumerate the lessons we learnt from this experiment and charts our way forward for the Spring semester.

# Table of Contents

# 2 Project Description

## 2.1 Refined Project Details

This project involves applying machine learning techniques in order to build an AI that can detect human emotions through a multitude of data modalities.

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

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.

## 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. A simplified representation of this is showcased below in fig 2.1:

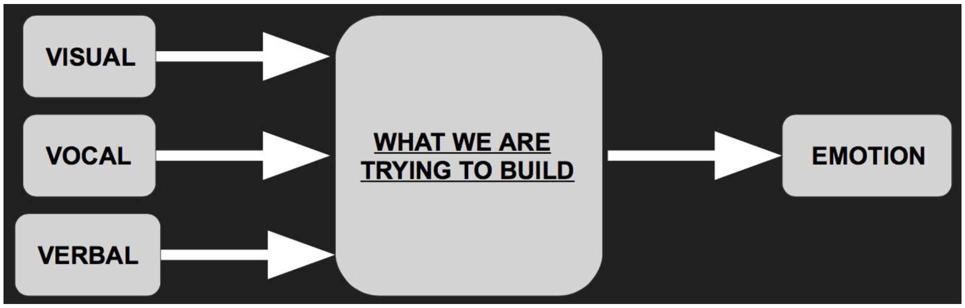


**Figure 2.1: High level representation of our goal**

The applications of this project are widespread. These include:

1. Enhancing social skills of current robot assistants( eg home robots, virtual assistants, etc)
2. 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. Honing social skills for the socially challenged, and coincidentally building a sense of understanding (and not repudiation) by the public towards people like this - This use case will be explored in greater detail below.

## 3. Use Case

Michael, illustrated in Fig 3.1, 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 his mind blocks, he is 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 3.1: A depressed Michael with his Olly [1]**

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 for catharsis. 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, illustrated in Fig 3.2, and has 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.

**Figure 3.2: A cheerful Michael with his Olly[2]**

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 by our team. 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.

It is also important to focus specifically on the ability for the Olly to read Michael's emotions. That very ability is what this project is all about. Multimodal emotion recognition is just a piece of the puzzle, (though a crucial one) that is required to engineer an agent that can interact with us at this level. This power of emotion recognition expands the scope of possible consequential actions, and so in the context of Olly, this multimodal emotion recognition system would be used in conjunction with a natural language model and speech technology system.

## 4. System-level requirements

The following is the breakdown of our project requirements. These requirements are categorized as mandatory (M) or desirable (D), as well as performance(P) and nonfunctional (N).

### 4.1 Functional requirements

Our project has two main parts. The first is the ability to detect emotions, the other is the ability to track users. The functional requirements of this project will be split up for each of these two parts.

**Table 4.1. Functional requirements details for emotion detection**

| ID | Title | Description |
|---|---|---|
| M.P1 | Shall detect 5 emotions from tri-modal data | The system will detect human emotional with an accuracy of up to 50% |
| M.P2 | Shall output emotion chart | The system will output a display of the emotion chart at 1 frame per second |

**Table 4.2. Functional requirements details for ability to track users**

| | | |
|---|---|---|
| M.P5 | Shall Track user | The camera will track the user 70% of the time that they are within the field of the view of the camera, with real-time speed |

### 4.2 Non-functional requirements

Table 4.3 enlists non-functional requirement details for the robot. We have no non-functional requirements that relate to emotion recognition, but rather to the face tracking hardware. We want to essentially emulate the product of our sponsor Emotech (Olly) which is a little robot like the Amazon Dot and and thereby the non-functional requirements were chosen.

**Table 4.3: Nonfunctional requirements**

| ID | Title | Description |
|---|---|---|
| D.N1 | Rests on tabletop | The robot that will serve as the physical casing for our instrument cluster shall rest on a 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. |
| D.N4 | Smaller than a microwave | The size of the robot itself shall be smaller than a microwave. |
| D.N5 | Less than 5kg | The robot shall weigh less than 5kg, (battery included) |

## 5. Functional Architecture

### 5.1 Block Diagram

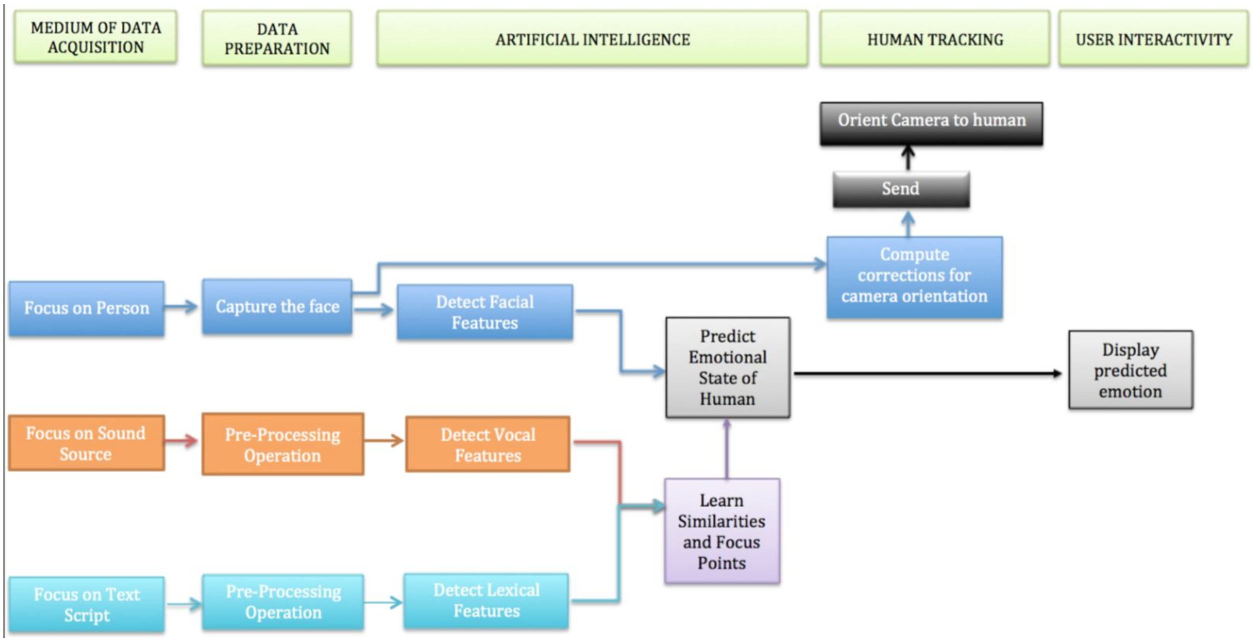


**Fig 5.1: Block diagram functional architecture**

### 5.2 Description of Functional Architecture

The desire is to be able to feed 3 different kinds of data streams through separate pipelines into our system and have an emotion recognition output, as shown in Fig 5.1. This raw data by itself shouldn't just be fed raw but rather preprocessed so as to better expose the neural networks to their interesting features. This phase is known as data preparation. After the data has been prepared, feature detection subsystems extract or learn features which present themselves in the multimodal emotional data. After the system has detected the features, it learns latent features that are shared by the audio and verbal data modalities. The learnt common features and focus points (that both modalities show are helpful in predicting emotion) are then fed along with visual features into a subsystem that "examines" these features in unison and uses them to predict emotion. This emotion is then displayed in way that is digestible to humans. Going back to the visual feature detection, since our project involves the ability both to detect emotions and track humans, the location of the facial features in the image is used as information to help orient the robot towards the human.

## 6. Cyber-Physical Architecture

### 6.1 Block Diagram

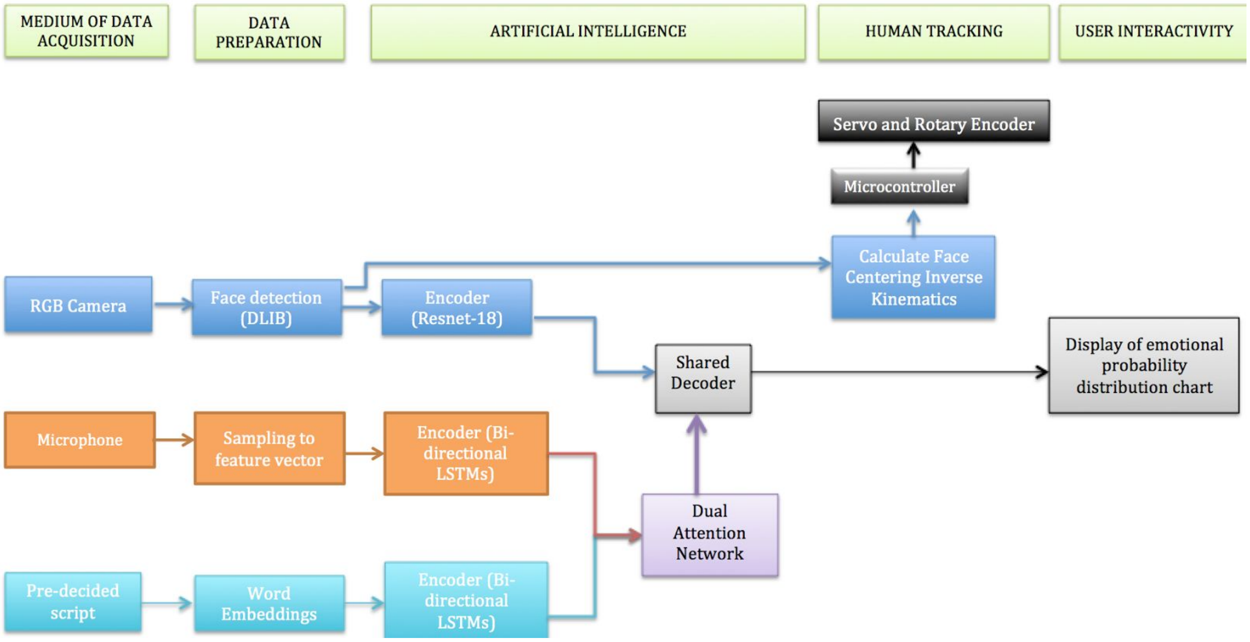


**Figure 6.1: Block diagram cyber physical architecture**

### 6.2 Description of cyber-physical architecture

Fig 6.1 illustrates the cyber-physical architecture of the system.

An RGB camera will be used to capture the raw video feed of the person who is speaking. This will be passed to the face detection program which will crop the face for each frame and this will serve as the visual input for the neural network. For the vocal modality, a microphone will be used to capture the raw waveform. This waveform will be sampled at a desired sampling rate frequency. During this conversion raw vectors will be extracted which will act as vocal input to the neural network. A pre-decided script will be used as raw input for the text modality. Each word will be converted to vectors to get word embeddings. These word embeddings will act as verbal input to the neural network.

After this there is an encoder for each modality. For the vision modality we are using the resnet-18 architecture. A bidirectional LSTM is used as an encoder for the vocal and the verbal modalities. The extracted features from the verbal and vocal modalities are fed into the dual attention network which will learn similarities and focus points between the two modes. The dual attention network gives as an output a shared vector which captures information between both the modalities. This is called a memory vector. This memory vector along with the extracted features from the vision modality is passed on to the shared decoder for final prediction of emotion. The shared decoder is an LSTM which learns temporal encoded features and makes predictions for each frame.

## 7. Current System Status

### 7.1 Fall Semester Requirements

### 7.1.1 Emotion detection accuracy of 50% across 5 emotions

All the various real-life emotions were compartmentalized into 5 distinct emotion neighborhoods/buckets for identification. The first bucket is happy/elated/excited. The second neighbourhood is sad/depressed/gloomy. The third cluster is upset/alarmed/fear/anger. The fourth one is calm/content/relaxed and the last one is neutral. The requirement was that the neural network should perform at an average accuracy of 50% on the test set of the SEMAINE dataset across these 5 emotional buckets.

### 7.1.2 Emotion detection rate of 1 frame per second.

The emotion recognition system should output an emotion for each frame in the test video at an average of 1 frame per second.

### 7.1.3 Face tracking accuracy of 70 %.

The bot should track a human’s face at a maximum distance of 1m from it. The human is supposed to move at a maximum speed of 10 cm/s left to right in front of it. The bot should be able to successfully keep the face of the human within the frame of its camera 70% of the length of the experiment.

## 7.2 Subsystem descriptions

### 7.2.1 Vision Subsystem

The underlying goal of the subsystem is to extract important visual features from raw video feed. Fig 7.1 shows the pipeline for achieving the same.

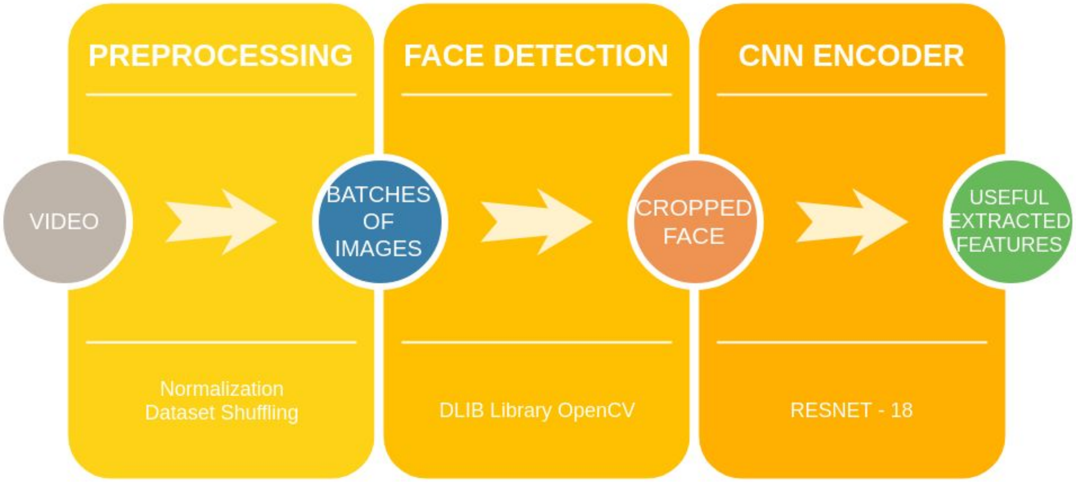


**Fig 7.1 : Overview of the Vision Subsystem Pipeline**

The preprocessing steps include dataset shuffling and image normalization for each video frame in the dataset. Faces are cropped from all of the images gathered in the above step using DLIB OpenCV. Batches of images are then fed into a CNN encoder which will extract features from them. The CNN encoder that is being used is the Resnet-18 architecture as shown in Fig 7.2

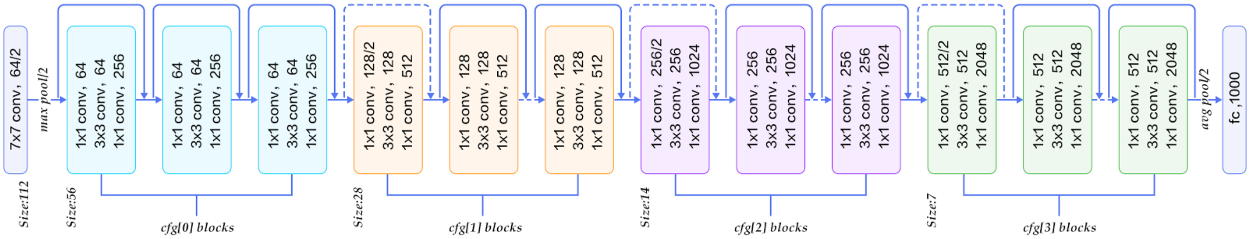
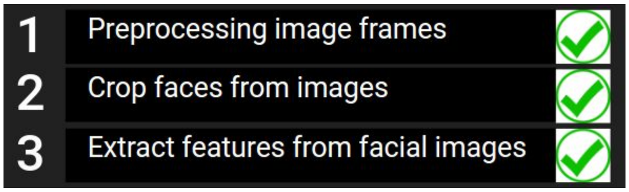


**Fig 7.2 : A general Resnet architecture[3]**

Our implementation of the Resnet Architecture comprised of a removed fully connected layer since we were only interested in feature extraction. To add, the output from the average pooling layer was connected to the multimodal shared decoder.

The Resnet-18 architecture was trained on the CREMA-D dataset which is a bimodal (visual + vocal) emotion recognition dataset. Once we trained the network, we further fine-tuned it on the SEMAINE dataset. This was done because SEMAINE focuses more on text and therefore the dataset consists of only 8 different faces. Therefore we wanted to diversify our training set and expose the network to a variety of faces.

Currently, the vision subsystem has been fully implemented, as shown in Fig 7.3. We have the preprocessing pipeline ready. Our vision encoder is also trained on tested on CREMA-D and SEMAINE.



1 Preprocessing image frames ✓
2 Crop faces from images ✓
3 Extract features from facial images ✓

**Fig 7.3: Status of the Vision Subsystem**

### 7.2.2 Vocal Subsystem

The goal of this subsystem is to extract important vocal features from raw waveform. Fig 7.4 shows a raw audio waveform from which features have to extracted.

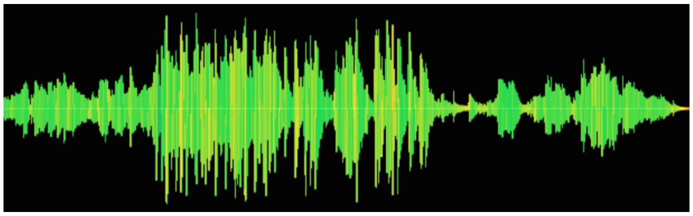


**Fig 7.4: Raw Audio Waveform**

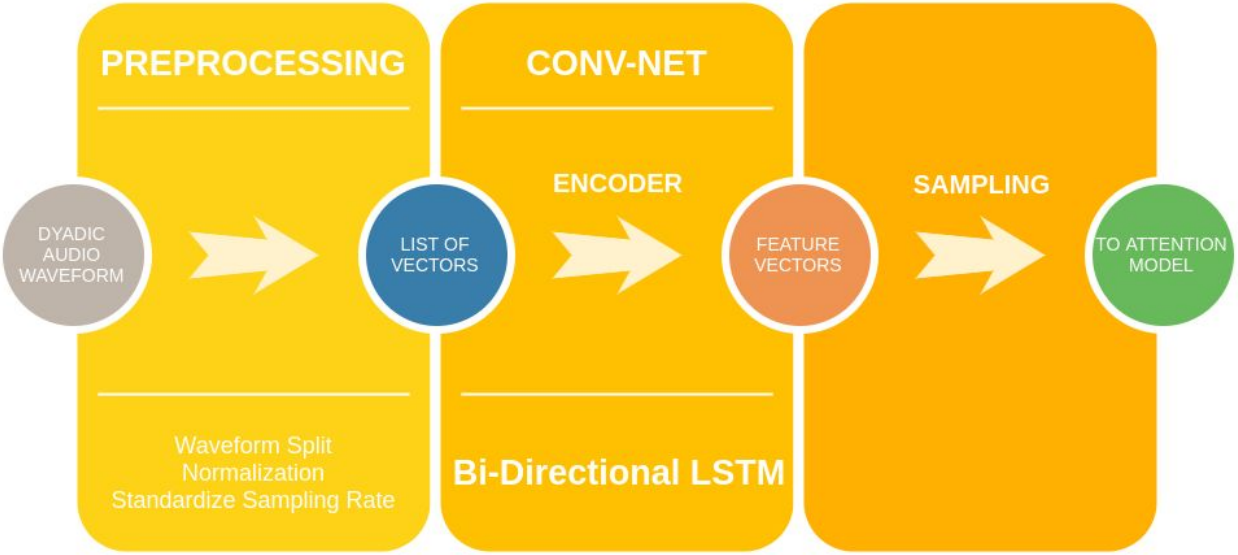


**Fig 7.5 : Vocal subsystem pipeline overview**

Fig 7.5 gives an overview of our vocal subsystem. The first step is splitting the waveform into a unit time interval according to our frame rate for the particular dataset. After that we sample the waveform at 16Khz frequency. This implies that we would have a 16000-d vector for each second of audio. We normalize the audio waveform by subtracting the mean and dividing it by the variance. After extracting multiple vectors at unit time intervals from the audio it is put through an encoder which will extract audio features. There are two options for the encoder which we are exploring. One of them is the CNN encoder which was implemented, trained and tested in the fall and the LSTM encoder will be tested and compared to the CNN encoder in the spring.

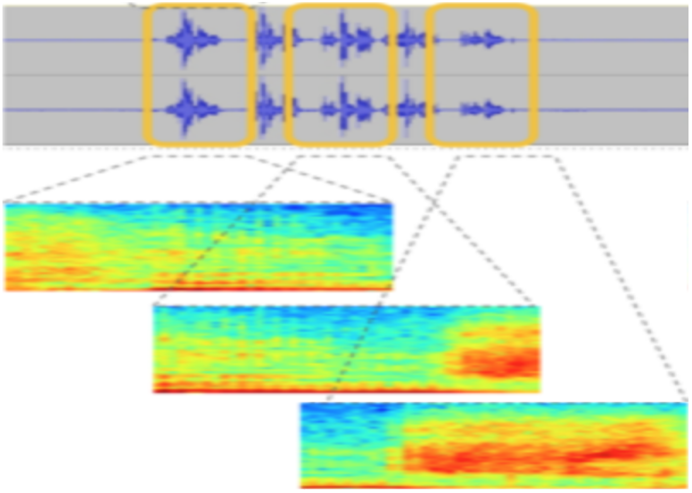
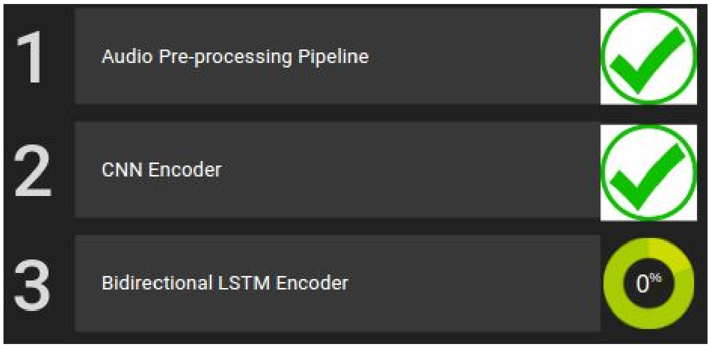


**Figure 7.6: Yellow convolutional filter boxes extracting temporal features on a waveform**

The CNN encoder has filters which span through time and extract temporal features as shown in Fig 7.6. The vocal encoder has been trained on the SEMAINE dataset. Currently, the audio preprocessing pipeline has been implemented and tested. The CNN encoder has been implemented , trained and tested. The Bidirectional LSTM Encoder will be implemented during the spring semester. Fig 7.7 shows the net status of vocal subsystem.

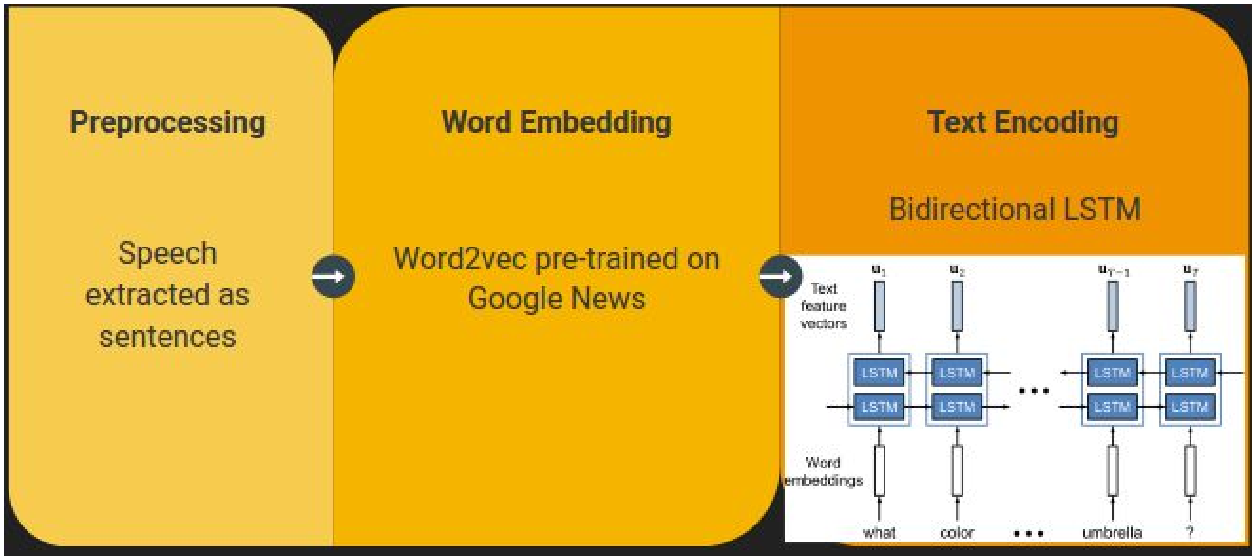


| | | |
|---|---|---|
| 1 | Audio Pre-processing Pipeline | ✓ |
| 2 | CNN Encoder | ✓ |
| 3 | Bidirectional LSTM Encoder | 0% |

**Fig 7.7: Status of the vocal subsystem**

### 7.2.3 Verbal Subsystem

The aim of this subsystem is to extract important learned features from raw text/transcript data. Fig 7.8 shows the pipeline for the verbal subsystem.



Preprocessing

Speech extracted as sentences

Word Embedding

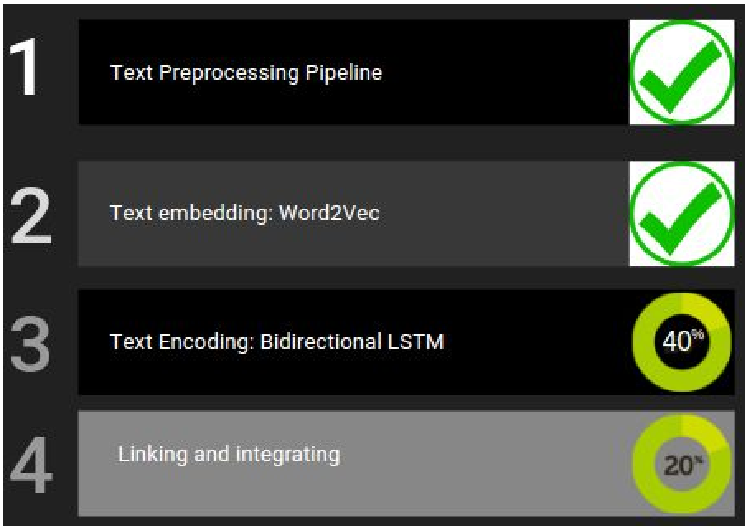
Word2vec pre-trained on Google News

Text Encoding

Bidirectional LSTM

**Fig 7.8: Overview of the verbal subsystem**

In the preprocessing stage the transcript is divided into words and each word is converted to a vector using word2vec which is pre-trained on Google News. These vectors are then passed to a bidirectional LSTM for extracting features from the text. The preprocessing and the word embedding for the verbal subsystem is complete. Apart from that, the implementation of the text encoder is encoder is complete, as shown in Fig 7.9. However it has to be integrated with the bimodal (visual+vocal) network that we have built in the fall, trained and tested .



1 Text Preprocessing Pipeline ✓

2 Text embedding: Word2Vec ✓

3 Text Encoding: Bidirectional LSTM 40%

4 Linking and integrating 20%

**Fig 7.9: Status of the verbal subsystem**

### 7.2.4 Attention Model

The goal of the attention model is to learn similarities between the vocal and the verbal modalities as well as focus points between them. This helps the overall model to learn emotions better and faster  by connecting representation of different modalities together. Fig 7.10 shows a high level overview of the attention model.

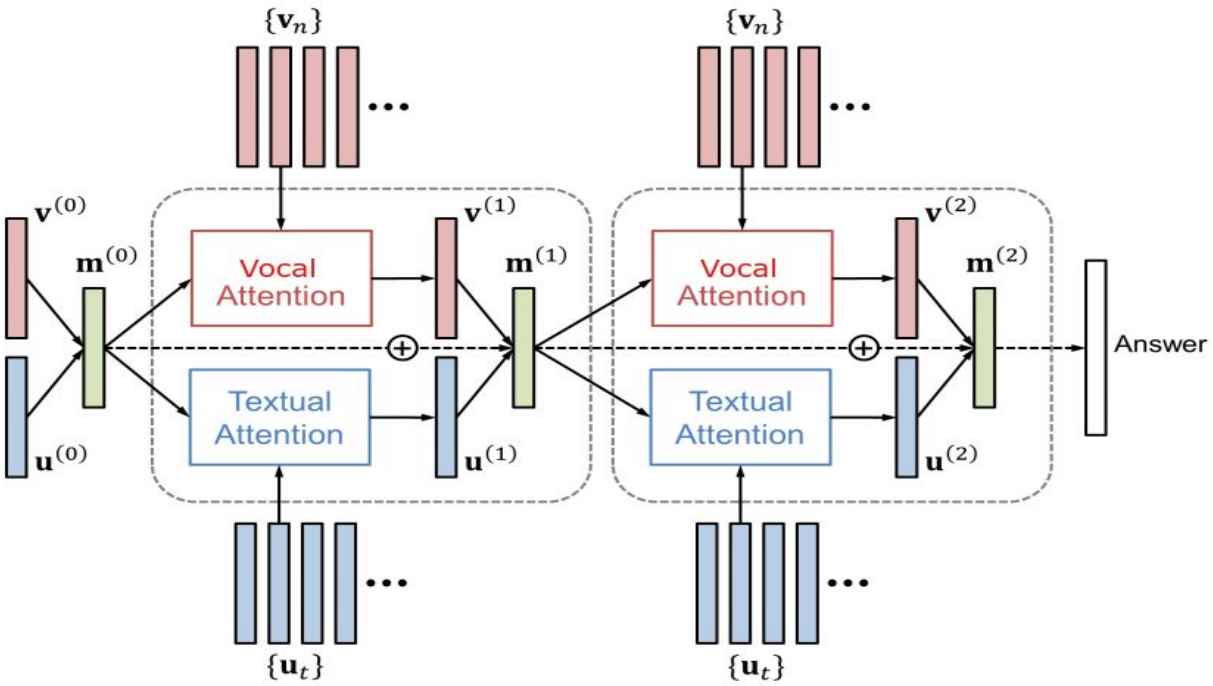


**Fig 7.10: Dual Attention Network[4]**

On a low level the attention model, shown in fig 7.10, is a simple feedforward neural network which takes in feature vectors from the vocal and the verbal encoders as well as a concatenated representation of a "memory" vector(denoted by m vector in the figure) which tries to lump all of the information learned by the attention model in the previous step. The above figure is a two step or a dual attention network. The feedforward neural network is repeated two times as shown by two bounding boxes. The memory vector at the end is passed to the multimodal shared decoder.

The implementation of the attention model, as shown in fig 7.11, is complete but it is yet to be integrated with the bimodal network and trained.

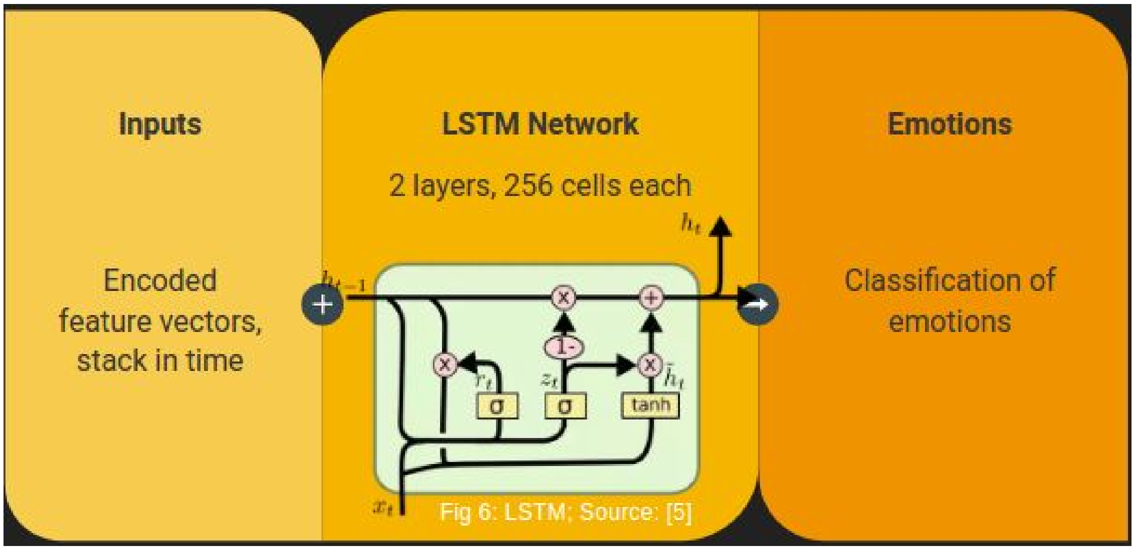
1 Implementation ✓

2 Training on datasets 0%

**Fig 7.11: Status of the Attention Model**

### 7.2.5 Multimodal Shared Decoder

The aim of this subsystem is to predict the final emotion given the feature vectors from vision subsystem and memory vector from the attention model as inputs. Fig 7.12 gives an overview of the subsystem.


Inputs

Encoded feature vectors, stack in time

LSTM Network

2 layers, 256 cells each

Emotions

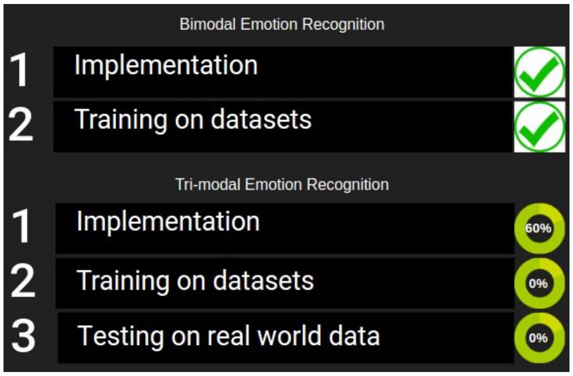
Classification of emotions

Fig 6: LSTM; Source: [5]

**Fig 7.12: Overview of Multimodal Shared Decoder**

After the feature vectors from the vision subsystem and the memory vector from the attention model have been concatenated, they are passed into a 2 layered LSTM with 256 cells each. The LSTM has its last hidden cell connected to a fully connected layer which is necessary for doing predictions. The network outputs 4 values of valence, arousal , power and expectation. These values are then interpreted into which one of the 5 emotional bucket it maps into.

For the fall a bimodal LSTM decoder was implemented which took as input a concatenated feature vector from the vision and the vocal encoder. For the spring the trimodal decoder will be trained and tested which takes the attention model and the visual features as inputs. However since the implementation of the decoder itself more or less remains the same except the input size it has been marked as partially completed. Status of the subsystem is shown in Fig 7.13


Bimodal Emotion Recognition

| | | |
|---|---|---|
| 1 | Implementation | ✓ |
| 2 | Training on datasets | ✓ |

Tri-modal Emotion Recognition

| | | |
|---|---|---|
| 1 | Implementation | 60% |
| 2 | Training on datasets | 0% |
| 3 | Testing on real world data | 0% |

**Fig 7.13 : Status of the Multimodal Shared Decoder Subsystem**

### 7.2.6 Robot Hardware

The goal of this subsystem is to track a human by detecting the movement of their face. The cropped face is then sent to the vision subsystem. The other objective is to record what the person is saying and relay that information to the network. Essentially the robot is just a physical manifestation of our AI agent. For the fall experiment, our goal was to track humans in 1 dimension, which thus required only 1 rotational actuator (a stepper) to orient its camera towards the human. This explains the simple design shown in fig 7.14, and in the spring, we will design a 2 DOF robot so that it can track a person in 2D both as they move across in front of it as well as towards and away from it. This a very small robot with a total height of 32cm from base of the stepper to the top of the webcam. Status of implementation is shown in fig 7.15.

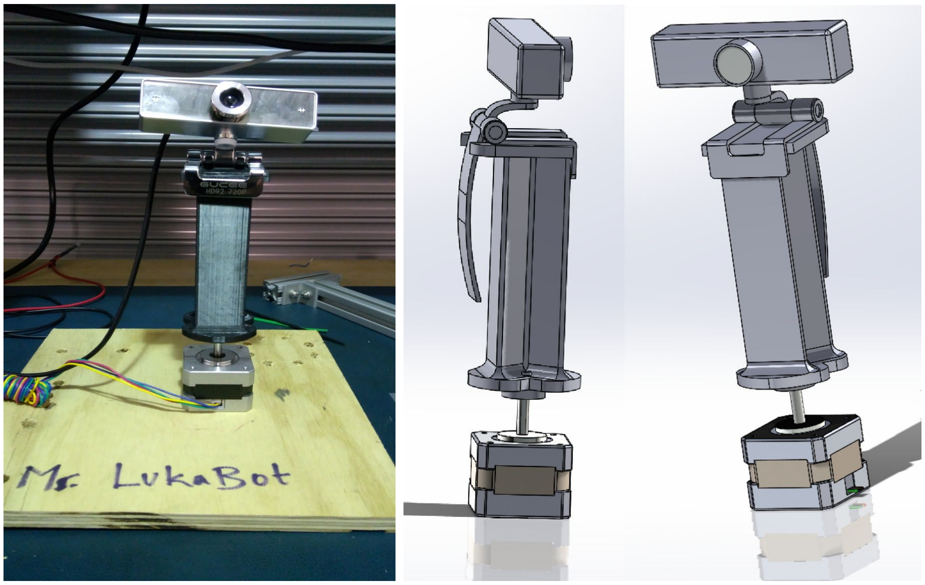
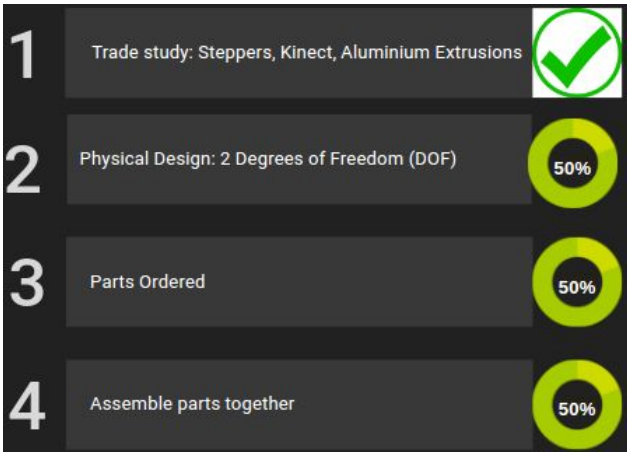


**Fig 7.14: LukaBot (assembled on the left, and designed in CAD on the right)**



1 Trade study: Steppers, Kinect, Aluminium Extrusions ✓

2 Physical Design: 2 Degrees of Freedom (DOF) 50%

3 Parts Ordered 50%

4 Assemble parts together 50%

**Fig 7.15: Status of the Robot Hardware Subsystem**

## 7.3 Modeling, Analysis and Testing

### 7.3.1 CAD Modeling of Face Tracker

The CAD model of face tracker is given in Fig 7.14 above. For SVE, this will have two actuators and two degrees of freedom.

### 7.3.2 Ablation Studies

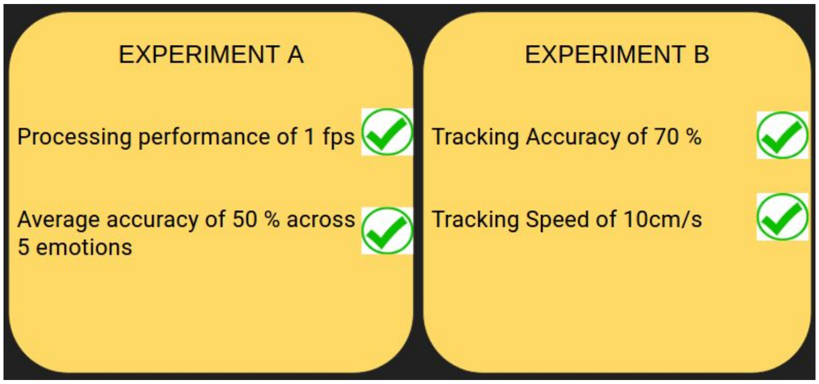
An ablation study refers to removing some part of the network architecture and seeing how that affects performance or systematically removing parts of the input to see which parts of the input are relevant to the networks output. We conducted this study and obtained the results below:

**Table 7.1 : Benchmark results of Ablation Studies**

| S.No | Network | Validation Accuracy (across 5 emotions) |
|---|---|---|
| 1 | Resnet-18 [pre trained on ImageNet] only | Poor convergence, 20 % |
| 2 | Resnet-18[no pre training] only | 24 % |
| 3 | Resnet-18[pretrained on CREMA-D] only | 32 % |
| 4 | Resnet-18 (from 3) + LSTM Decoder | 42 % |
| 5 | Resnet-18(from 3) + Vocal CNN encoder only | 40 % |
| 6 | Resnet-18(from 3) + Vocal CNN encoder + LSTM | 49.5 % |

## 7.4 Fall Validation Experiment Evaluation

The metrics of FVE evaluation are shown in Fig 7.16.



EXPERIMENT A

Processing performance of 1 fps ✓

Average accuracy of 50 % across 5 emotions ✓

EXPERIMENT B

Tracking Accuracy of 70 % ✓

Tracking Speed of 10cm/s ✓

**Fig 7.16 : Overview of FVE Evaluation**

### 7.4.1 Experiment A

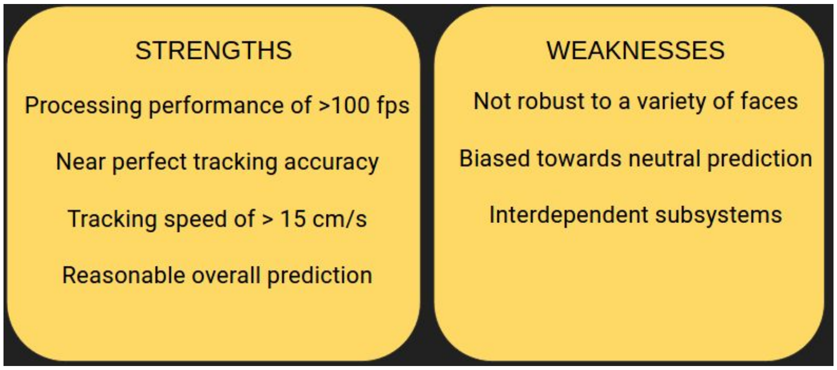
The goal of experiment A was to test the emotion recognition system on the test set of SEMAINE dataset. The success criterion included a processing performance of 1 frame per second and an average accuracy of 50% across 5 emotions across all 8 test videos. Both of those criterion were met. The processing performance of our system was more than 100 fps and our average emotion recognition accuracy across 3 different methods of decoding emotions using valence, arousal , power and expectation was 50.6%. Therefore we successfully passed both our requirements for experiment A during the FVE and the FVE Encore.

### 7.4.2 Experiment B

The goal of experiment B was to track a person's face whilst they moved at a maximum distance of 1 metre from the camera. The success criterion stated that the person should be successfully tracked 70% of the length of the experiment whilst he is moving at a maximum speed of 10 cm/sec left to right in front of the camera. We were able to achieve 100% tracking accuracy and a tracking speed of ~15 cm/sec during the FVE and the FVE Encore.

## 7.5 Strengths and Weaknesses

Fig 7.17 enlists strengths and weaknesses of our system.



STRENGTHS

Processing performance of >100 fps

Near perfect tracking accuracy

Tracking speed of > 15 cm/s

Reasonable overall prediction

WEAKNESSES

Not robust to a variety of faces

Biased towards neutral prediction

Interdependent subsystems

**Fig 7.17: Overview of strengths and weakness**

### 7.5.1 Strengths

- Processing Performance
  The speed of the network at which it predicts is greater than 100 fps which is faster than 99% of movie recordings/ mobile recordings which are recorded at ~ 24 fps. This means that the system has the ability to output emotions in real time without lag.
- Near Perfect Tracking Accuracy
  During experiments conducted during the FVE and the FVE encore we were able to achieve 100 % accuracy by keeping the human's face within the field of view of the camera at all times.
- Tracking speed
  We are able to achieve successful tracking even when the human walks at a pace greater than 15 cm/sec.
- Overall Emotion Prediction

We were able to achieve reasonable overall prediction for the emotion recognition subtask. The system is able to recognize all emotions particularly when they are clearly not subtle.

### 7.5.2 Weaknesses

- Diversity of faces

  The vision encoder is trained on 100 faces. This is clearly not enough during the spring when it will be tested against real world humans.
- Prediction Bias

  The network is currently biased towards predicting neutral emotion. The reason for this is that many videos in the dataset were neutral, which is probably due to the fact that most interactions in the real world are also fairly neutral.
- Interdependent Subsystems

  The network has all subsystems interconnected except the hardware. This makes it harder to know why the model works or why it doesn't work or why it behaves the way it does easily similar to many deep learning systems. It's harder to debug huge end to end networks. Ablation studies are time consuming to train and test and they too have some modifications in individual subsystems put into them to make it work.

# 8. Work Breakdown Structure

## 8.1 Work Breakdown Structure

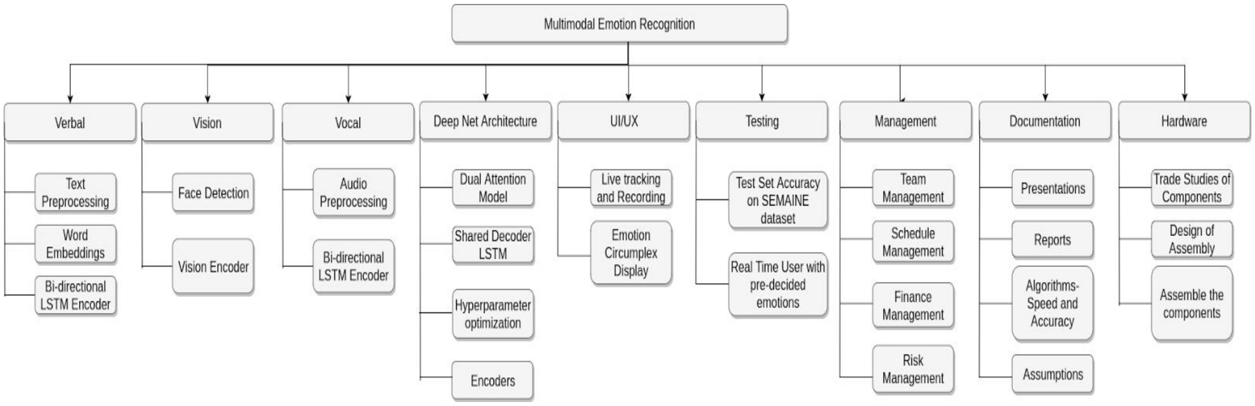
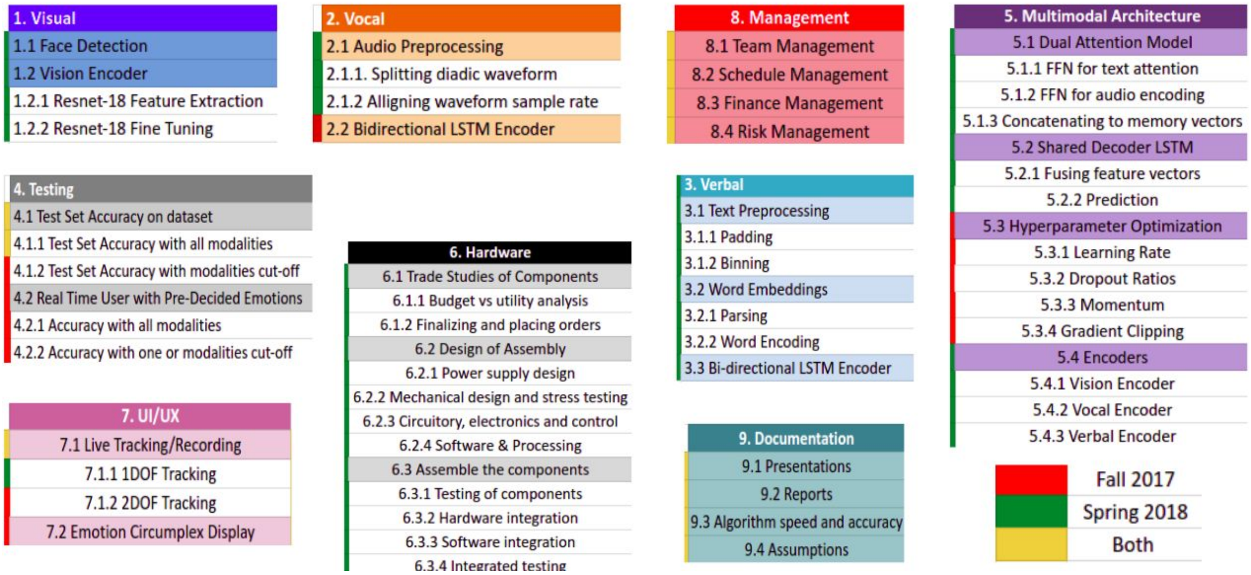


**Fig 8.1: Three-level Work Breakdown Structure**

As shown in the figure 8.1, we have 4 work packages on the software end one corresponding to each of the modes and one for the trimodal system. We have one associated to the robot hardware, one for UI/UX, one for testing/training, and two for project management and documentation.



**1. Visual**
- 1.1 Face Detection
- 1.2 Vision Encoder
- 1.2.1 Resnet-18 Feature Extraction
- 1.2.2 Resnet-18 Fine Tuning

**2. Vocal**
- 2.1 Audio Preprocessing
- 2.1.1. Splitting diadic waveform
- 2.1.2 Alligning waveform sample rate
- 2.2 Bidirectional LSTM Encoder

**8. Management**
- 8.1 Team Management
- 8.2 Schedule Management
- 8.3 Finance Management
- 8.4 Risk Management

**5. Multimodal Architecture**
- 5.1 Dual Attention Model
- 5.1.1 FFN for text attention
- 5.1.2 FFN for audio encoding
- 5.1.3 Concatenating to memory vectors
- 5.2 Shared Decoder LSTM
- 5.2.1 Fusing feature vectors
- 5.2.2 Prediction
- 5.3 Hyperparameter Optimization
- 5.3.1 Learning Rate
- 5.3.2 Dropout Ratios
- 5.3.3 Momentum
- 5.3.4 Gradient Clipping
- 5.4 Encoders
- 5.4.1 Vision Encoder
- 5.4.2 Vocal Encoder
- 5.4.3 Verbal Encoder

**4. Testing**
- 4.1 Test Set Accuracy on dataset
- 4.1.1 Test Set Accuracy with all modalities
- 4.1.2 Test Set Accuracy with modalities cut-off
- 4.2 Real Time User with Pre-Decided Emotions
- 4.2.1 Accuracy with all modalities
- 4.2.2 Accuracy with one or modalities cut-off

**6. Hardware**
- 6.1 Trade Studies of Components
- 6.1.1 Budget vs utility analysis
- 6.1.2 Finalizing and placing orders
- 6.2 Design of Assembly
- 6.2.1 Power supply design
- 6.2.2 Mechanical design and stress testing
- 6.2.3 Circuitory, electronics and control
- 6.2.4 Software & Processing
- 6.3 Assemble the components
- 6.3.1 Testing of components
- 6.3.2 Hardware integration
- 6.3.3 Software integration
- 6.3.4 Integrated testing

**3. Verbal**
- 3.1 Text Preprocessing
- 3.1.1 Padding
- 3.1.2 Binning
- 3.2 Word Embeddings
- 3.2.1 Parsing
- 3.2.2 Word Encoding
- 3.3 Bi-directional LSTM Encoder

**7. UI/UX**
- 7.1 Live Tracking/Recording
- 7.1.1 1DOF Tracking
- 7.1.2 2DOF Tracking
- 7.2 Emotion Circumplex Display

**9. Documentation**
- 9.1 Presentations
- 9.2 Reports
- 9.3 Algorithm speed and accuracy
- 9.4 Assumptions

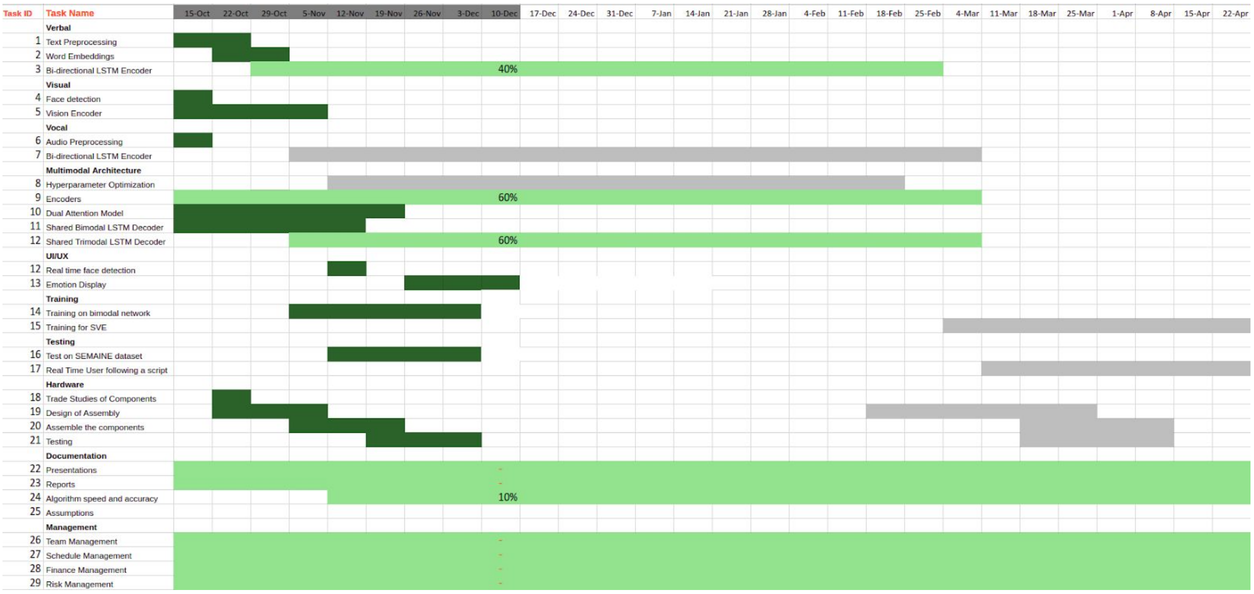
| | |
|---|---|
| 🟥 | Fall 2017 |
| 🟩 | Spring 2018 |
| 🟨 | Both |

**Fig 8.2: Work packages**

Figure 8.2 shows the work pages in a more detailed manner annotated by semesters they're scheduled to be finished in.

## 8.2 Schedule

### 8.2.1 Weekly Schedule



| Task ID | Task Name |
|---|---|
| | **Verbal** |
| 1 | Text Preprocessing |
| 2 | Word Embeddings |
| 3 | Bi-directional LSTM Encoder |
| | **Visual** |
| 4 | Face detection |
| 5 | Vision Encoder |
| | **Vocal** |
| 6 | Audio Preprocessing |
| 7 | Bi-directional LSTM Encoder |
| | **Multimodal Architecture** |
| 8 | Hyperparameter Optimization |
| 9 | Encoders |
| 10 | Dual Attention Model |
| 11 | Shared Bimodal LSTM Decoder |
| 12 | Shared Trimodal LSTM Decoder |
| | **UI/UX** |
| 12 | Real time face detection |
| 13 | Emotion Display |
| | **Training** |
| 14 | Training on bimodal network |
| 15 | Training for SVE |
| | **Testing** |
| 16 | Test on SEMAINE dataset |
| 17 | Real Time User following a script |
| | **Hardware** |
| 18 | Trade Studies of Components |
| 19 | Design of Assembly |
| 20 | Assemble the components |
| 21 | Testing |
| | **Documentation** |
| 22 | Presentations |
| 23 | Reports |
| 24 | Algorithm speed and accuracy |
| 25 | Assumptions |
| | **Management** |
| 26 | Team Management |
| 27 | Schedule Management |
| 28 | Finance Management |
| 29 | Risk Management |

**Fig 8.3 Gantt Chart**

Weekly schedule of activities have been documented and are being tracked. A current snapshot can be seen in Fig 8.3.

### 8.2.2 Briefing

**What are the major system development milestones in the remaining schedule?**

Major milestones in development are as below:

- Integrating text modality to get tri-modal network
- Integrating attention between text and vocal modalities
- Installing filters for vocal mode and noise filters for ratings
- 2D face tracking and integration with network
- Training and testing on real world data, exposing to multiple existing datasets, tuning hyperparameters, etc

**Are you behind, ahead of, or on schedule? If behind, how will you catch up?**

We're behind schedule since we'd underestimated the difficulty of the task at hand. We plan to catch up by gaining lost ground in winter and downscoping requirements to focus on the most important problems.

## 8.3 Test Plan

### 8.3.1 Milestones for Spring Semester

**Table 8.1: Milestones for Progress Reviews**

| PR Sl.No. | Timing | Goals |
|---|---|---|
| PR 7 | Late January | ● Achieve face tracking in two dimensions<br>● Implement and test text only network |
| PR 8 | Mid-February | ● Finalize trimodal recognition system after integrating text<br>● Start training on datasets |
| PR 9 | Late February | ● Testing trimodal network<br>● Adding filters to the network |
| PR 10 | Mid-March | ● Test tri-modal network on real-time users<br>● Fine tuning and tweaking for better results |
| PR 11 | Early April | ● Integrate hardware & recognition system<br>● Fine tuning and tweaking for better results |
| PR 12 | Mid-April | ● Tune the system and prepare for SVE |

### 8.3.2 Spring Validation Experiment

**Experiment A: Face Tracking System**
Location: Newell-Simon Hall, B floor
Test Conditions:

- Indoor room with lighting conditions from 4000 lux to 5000 lux
- Single active subject pacing at a maximum speed of
  - 10 cm/sec (Left - Right)
  - 3 cm/sec (Forward - Backward)

Experiment length : 1 min
Expected Result: System will track face with following metrics
Metrics:

- Face must be kept within the frame for 70 % of the experiment length

**Experiment B: Emotion Recognition System**
Location: Newell-Simon Hall, B floor
Test Conditions

- Indoor room with lighting conditions from 4000 lux to 5000 lux
- Subject acts out a pre-decided script in English from a distance of 30-45 cm in front of the camera.
- Single active subject, face may move at pace of
  - 5 cm/sec (Left - Right)
  - 3 cm/sec (Forward - Backward)

Expected Result: System will detect emotion with following performance metrics

- Speed: 1 frame/s
- Accuracy: 50% across 5 emotions
- Tracking: Keep face within the frame for 70% of experiment time

## 8.4 Budget

### 8.4.1 Refined Parts List

**Table 8.2: Refined parts list and costs incurred**

| Sl. No. | Part Name | Purpose | Quantity | Part Specification | Unit Price | Total Price |
|---|---|---|---|---|---|---|
| 1. | Microphones | Audio recording | 2 | Pyle Pro USB | $28.99 | $57.98 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2. | Servo Motors | Actuator for face tracking | 5 | Micro Servo - High Powered, High Torque Metal Gear | $11.95 | $59.75 |
| 3. | Camera | Visual input | 2 | Webcam 720P HD | $19.99 | $39.98 |
| 4. | Contingency camera | With attached microphone | 2 | Ausdom Full HD 1080p r | $22.99 | $45.98 |
| 5. | Contingency servo motors | Higher torque | 2 | HS-805MG Servo | $59.99 | $119.98 |
| 6. | Mounting hubs | Hardware Assembly | 4 | Universal Mounting Hub 5mm | $7.49 | $29.96 |
| 7. | Screws | Hardware Assembly | 1 | Screws 2-56 | $12.60 | $12.60 |
| 8. | Nuts | Hardware Assembly | 1 | 2-56 Nuts | $8.00 | $8.00 |
| 9. | Contingency screws | Hardware Assembly | 1 | Steel Pan Head Machine Screw, (Pack of 100) | $6.17 | $6.17 |
| 10. | Contingency Nuts | Hardware Assembly | 1 | Hillman 140009 Zinc Hex, 4-40, 100-Pack | $4.25 | $4.25 |
| 11. | Hard drive | Data storage | 1 | WD 6TB Elements | $189.99 | $189.99 |
| 12. | LEDs | PCB | 40 | LED Vf: 2.2V | $1.22 | $48.80 |
| 13. | Zener diode | Reverse voltage protection | 40 | Zener Diode- 18V | $0.24 | $9.60 |
| 14. | Schottky Diode | Reverse current protection | 40 | Schottky Diode-2.5A | $0.47 | $18.80 |
| 15. | Laptop mouse | Infrastructure | 2 | WiRed USB Optical Mouse | $5.44 | $10.88 |
| - | - | - | - | - | Total | $661.84 |
| 16. | Older parts | Stepper motors, kinects | - | Scratched under revised plan | - | $706.48 |

| | | | | | | |
|---|---|---|---|---|---|---|
| . | Shipping and Handling | - | - | - | - | $10.0 |
| | | | | | Total | $1,368.32 |

### 8.4.2 Budget Summary

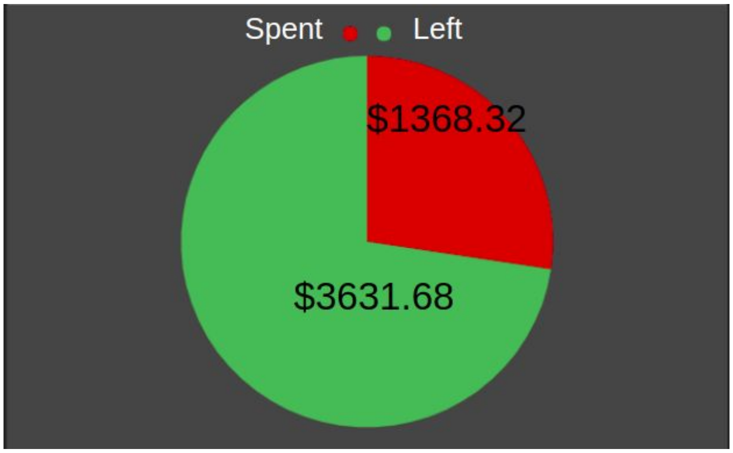


**Fig 8.4 Budget status**

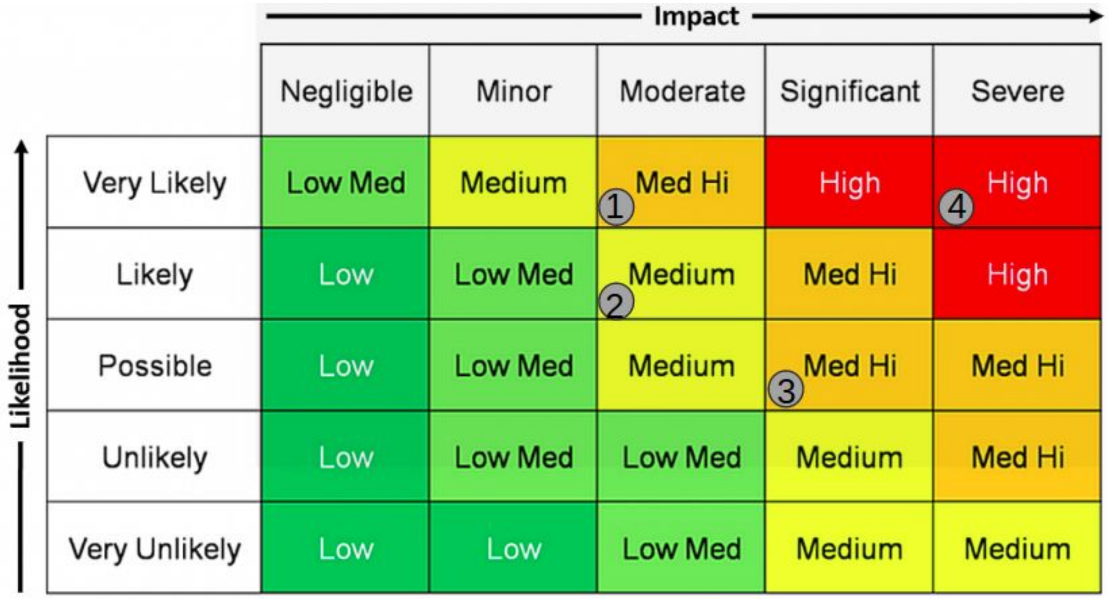
Our total budget is $5000.00. We have spent 27.37% of our total budget, as shown in figure 8.4. Currently our major expenses are the servo motors and hard drive. We have spent about 14% of our cost on stepper motors and microsoft kinects that we will not be using in the revised hardware design. Returning this will help us replenish our budget once again.

## 8.5 Risk management

### 8.5.1 Risk Identification

We have identified the following major risks and are tracking them closely:

1. Lack of diversity in datasets
2. Delay due to high training time
3. Hardware breakdown
4. Low accuracy

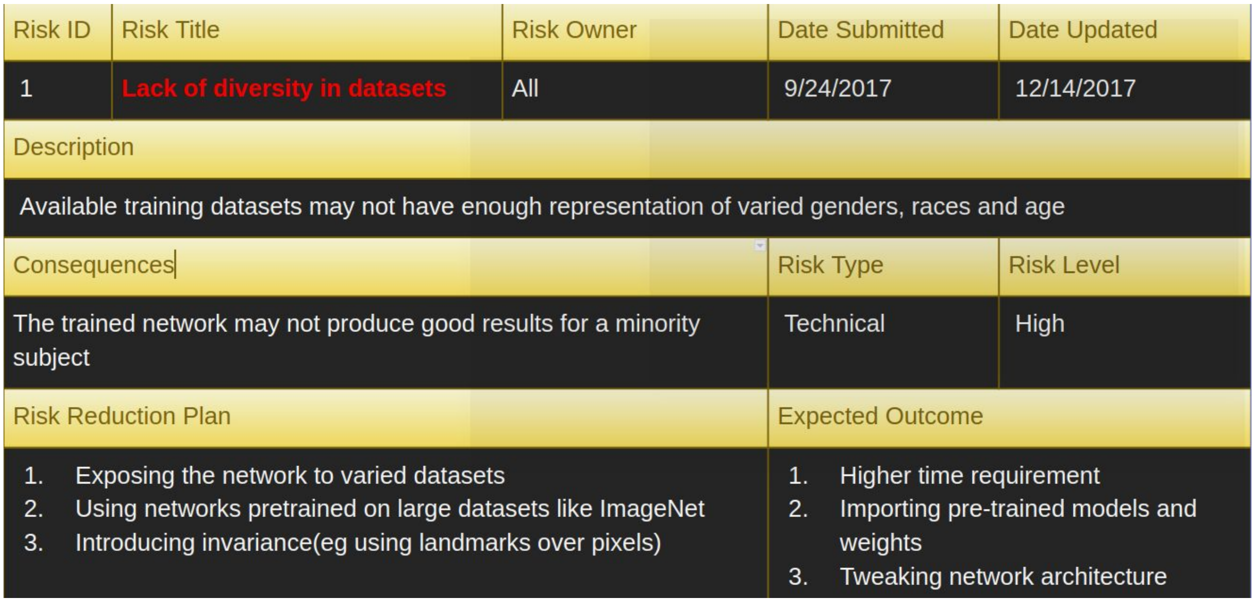


| Likelihood \ Impact | Negligible | Minor | Moderate | Significant | Severe |
|---|---|---|---|---|---|
| Very Likely | Low Med | Medium | 1 Med Hi | High | 4 High |
| Likely | Low | Low Med | 2 Medium | Med Hi | High |
| Possible | Low | Low Med | Medium | 3 Med Hi | Med Hi |
| Unlikely | Low | Low Med | Low Med | Medium | Med Hi |
| Very Unlikely | Low | Low | Low Med | Medium | Medium |

**Fig 8.5: Risk impact vs likelihood table for major risks**

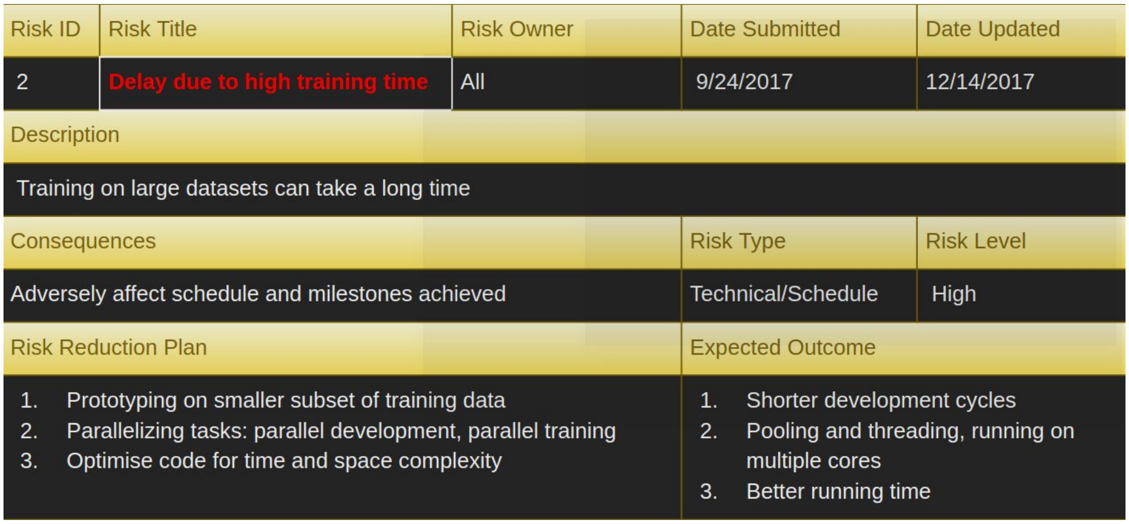
Fig 8.5 shows the Impact-likelihood ratings of different risks

### 8.5.2 Risk Mitigation



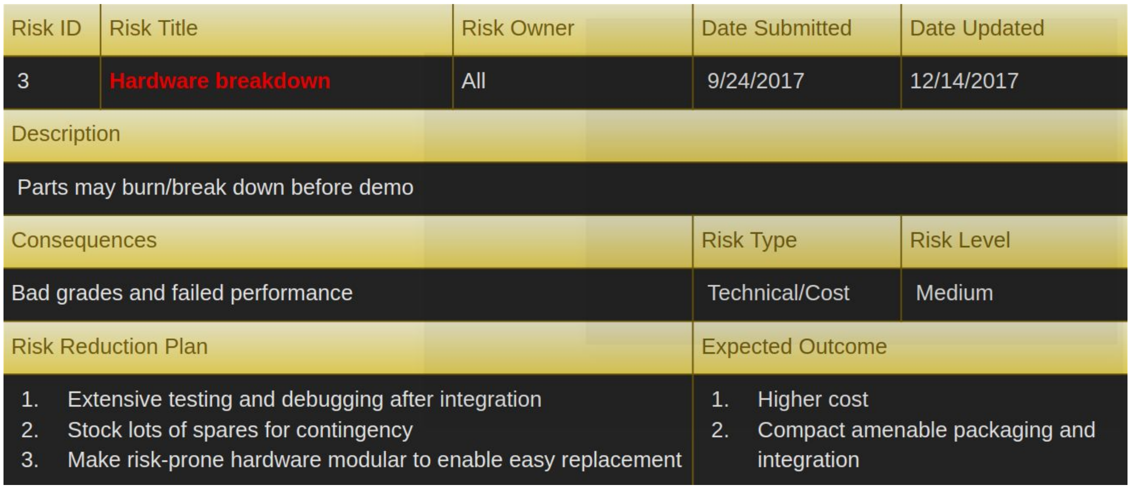
| Risk ID | Risk Title | Risk Owner | Date Submitted | Date Updated |
|---|---|---|---|---|
| 1 | Lack of diversity in datasets | All | 9/24/2017 | 12/14/2017 |
| Description | | | | |
| Available training datasets may not have enough representation of varied genders, races and age | | | | |
| Consequences | | | Risk Type | Risk Level |
| The trained network may not produce good results for a minority subject | | | Technical | High |
| Risk Reduction Plan | | | Expected Outcome | |
| 1. Exposing the network to varied datasets<br>2. Using networks pretrained on large datasets like ImageNet<br>3. Introducing invariance(eg using landmarks over pixels) | | | 1. Higher time requirement<br>2. Importing pre-trained models and weights<br>3. Tweaking network architecture | |

**Fig 8.6: Risk card for Risk ID 1**



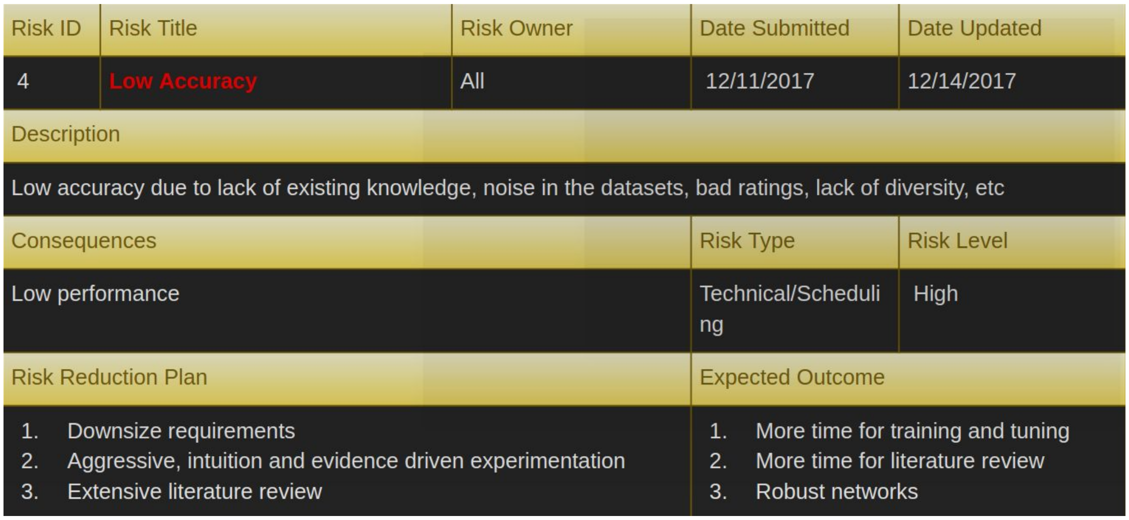
| Risk ID | Risk Title | Risk Owner | Date Submitted | Date Updated |
|---|---|---|---|---|
| 2 | Delay due to high training time | All | 9/24/2017 | 12/14/2017 |
| Description | | | | |
| Training on large datasets can take a long time | | | | |
| Consequences | | | Risk Type | Risk Level |
| Adversely affect schedule and milestones achieved | | | Technical/Schedule | High |
| Risk Reduction Plan | | | Expected Outcome | |
| 1. Prototyping on smaller subset of training data<br>2. Parallelizing tasks: parallel development, parallel training<br>3. Optimise code for time and space complexity | | | 1. Shorter development cycles<br>2. Pooling and threading, running on multiple cores<br>3. Better running time | |

Fig 8.4: Risk card for Risk ID 2



| Risk ID | Risk Title | Risk Owner | Date Submitted | Date Updated |
|---|---|---|---|---|
| 3 | Hardware breakdown | All | 9/24/2017 | 12/14/2017 |
| Description | | | | |
| Parts may burn/break down before demo | | | | |
| Consequences | | | Risk Type | Risk Level |
| Bad grades and failed performance | | | Technical/Cost | Medium |
| Risk Reduction Plan | | | Expected Outcome | |
| 1. Extensive testing and debugging after integration<br>2. Stock lots of spares for contingency<br>3. Make risk-prone hardware modular to enable easy replacement | | | 1. Higher cost<br>2. Compact amenable packaging and integration | |

Fig 8.8 Risk card for Risk ID 3



| Risk ID | Risk Title | Risk Owner | Date Submitted | Date Updated |
|---|---|---|---|---|
| 4 | Low Accuracy | All | 12/11/2017 | 12/14/2017 |
| Description | | | | |
| Low accuracy due to lack of existing knowledge, noise in the datasets, bad ratings, lack of diversity, etc | | | | |
| Consequences | | | Risk Type | Risk Level |
| Low performance | | | Technical/Scheduling | High |
| Risk Reduction Plan | | | Expected Outcome | |
| 1. Downsize requirements<br>2. Aggressive, intuition and evidence driven experimentation<br>3. Extensive literature review | | | 1. More time for training and tuning<br>2. More time for literature review<br>3. Robust networks | |

Fig 8.9 Risk card for Risk ID 4

Risk mitigation cards for the four major identified risks have been given in figures 8.6, 8.7, 8.8 and 8.9.

# 9. Conclusions

## 9.1 Key lessons learned during the fall semester

### 9.1.1 Low performance of state of the art multimodal sentiment analysis

In a very recent (July 2017) paper published by Dr. LP Morency and his students from CMU LTI on multimodal sentiment analysis across 5 sentiments the accuracy achieved on CMU-MOSI was ~42 %. Our project also deals with 5 emotions and sentiment represents the valence dimension of emotion very closely. Emotions are defined not only by valence(sentiment) but also by arousal. Since the accuracy on a theoretical dataset which is made under controlled conditions hovers around 40% we have lowered our performance requirements for the spring semester on real-world conditions to 50%.

### 9.1.2 Poor convergence on end to end training

It was observed that we couldn't get the end to end network to converge properly. We feel that the reason for this is probably vanishing gradients and it was very hard to find a global optima. We solved this by training the encoders separately and then adding the decoder and fine-tuning the entire network again. This maybe our future plan of action if we find similar behaviour on other datasets as well.

### 9.1.3 Training on multiple datasets is vital

Since the final product has to perform on real world conditions it is critical to train the network on as many datasets as possible so as to remove the bias of one or two datasets on the final model.

### 9.1.4 Data preprocessing is the most time consuming task

Every dataset has its own surprises and is rarely ever plug and play a preprocessing pipeline. For instance there are frames in which the face of the user goes outside the frame while laughing or while being excited. The number of raters for each video is different, some of the raters stopped rating after half of the video, the audio waveform is recorded at a different frequency than what is required by us and so on. It is not possible to write a script without knowing how each dataset is structured. In some worst cases, it can be identified and weeded out by close manual examination only.

### 9.1.5 On subtle examples, text is king.

For the fall we had a bimodal network which only used the visual and the vocal modality. While going through the validation and the test videos to examine the performance of the network as a human we found that transcript was important and it would be very difficult for even humans to differentiate on subtle cases the emotional state of the person.

### 9.1.6 Considering using servos instead of steppers

We originally started the project by using Steppers however their slow response rate and the somewhat jittery, stochastic tracking they have provided has pushed us to use Servos instead. Servos offer a quicker response rate, a smoother, more rapid ability to turn and a simpler method of programming. Servos can also be directly power by the Arduino which means that we can stop using a Rechargeable Turnigy LIPO battery to power the motors. This is a bonus for our Risk Mitigation Strategy as the success of our SVE will be less dependent on batteries working well, and the PCB will be simpler and smaller, meaning that there will be less connections that run the risk of shorting as we will be using less components on the circuit board.

## 9.2 Key Activities during the spring semester

### 9.2.1 Incorporating text modality

We have implemented and tested a bimodal network during the fall semester. For the spring semester we plan to add text modality to further improve the performance of the system.

### 9.2.2 Incorporating attention model

The attention model will help the network to converge faster by focusing on the "important" parts. This key item is therefore also on our ticket for the spring semester.

### 9.2.3 Adding frequency filters for vocal waveform processing

Currently, vocal convolutional filters extract features only across time. This will be extended to extract features across frequency also in the spring. This will be done using filter banks. The motivation behind this is to learn information along both dimensions of a waveform rather than only along time.

### 9.2.4 Fighting dataset bias

An ideal dataset should have an equal proportion of all emotions. But many emotional datasets are labeled by emotional dimensions of valence and arousal. Therefore it is harder to define bias towards a particular emotion in such a case. In the spring before training the network, emotion classification will be performed by converting those dimensions and an equal proportion of all emotions will be there in the training set to ensure that the network doesn't have a bias towards a particular emotion. This might involve data augmentation and/or discarding some data.

### 9.2.5 Face tracking in 2-D

Currently, we have in one place a 1-D face tracker which works for people moving left to right in front of the robot. In the spring the robot should be able to track people moving from left to right as well as people moving forward and backward simultaneously.

### 9.2.6 Modifying for real world conditions

For the spring semester, we will be testing the network on real world conditions as opposed to controlled conditions on which datasets are built. Based on the performance of the network we might try to emulate similar noisy features on the datasets and finetune the network.

This will not only serve as data augmentation but also will help the network to improve real-world performance.

## 10. References

[1]https://g68qpy3g1w-flywheel.netdna-ssl.com/wp-content/uploads/2014/07/depressed-worker-e1405446545412.jpg
[2]https://thumbs.dreamstime.com/z/happy-office-worker-desk-vector-illustration-35476920.jpg
[3] http://paddlepaddle.org/docs/develop/book/03.image_classification/index.html
[4] Nam H, Ha JW, Kim j, "Dual Attention Networks for Multimodal Reasoning and Matching", arXiv:1611.00471, https://arxiv.org/abs/1611.00471
[5] http://colah.github.io/posts/2015-08-Understanding-LSTMs