



16681 - MRSD Project Course

Final Report

9th May 2018

Team D: Deeply Emotional

Members:

Ritwik Das

Luka Eerens

Keerthana P G

Sponsor:

Emotech Ltd



1. Prefatory Information

This project 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. Our work demonstrates the use of multimodal fusion techniques in the field of emotion recognition. This final report nails down the refined overview of our use case, documents our system design and cyber physical architectures various aspects of project management. Additionally, we enumerate the lessons we learnt from this project and discuss possible future work on top of this project.

Table of Contents

1. Prefatory Information	1
2. Project Description	1
2.1 Refined Project Details	1
2.2 Project Goals	1
3. Use Case	2
4. System-level requirements	4
4.1 Functional requirements	4
4.2 Non-functional requirements	4
5. Functional Architecture	5
5.1 Block Diagram	5
5.2 Description of Functional Architecture	5
6. System Level Trade Studies	5
6.1 Joint Representation versus Coordinated Representation:	5
6.2 Early Fusion Versus Late Fusion:	6
6.3 Action Unit Features Versus Dense CNN Features:	7
6.4 Steppers Versus Servos:	8
7. Cyber-Physical Architecture	8
7.1 Block Diagram	8
8. System Description and Evaluation	9
8.2 Subsystem descriptions	10
8.2.1 Vision Subsystem	10
8.2.2 Vocal Subsystem	12
8.2.3 Verbal Subsystem	13
8.2.4 Attention Model	14
8.2.5 UI	15
8.2.6 Robot Hardware	17
8.3 Modeling, Analysis and Testing	18
8.3.1 CAD Modeling of Face Tracker	18

8.3.2 Experimental Results	18
8.4 SVE Requirements	20
8.4.1 Emotion detection accuracy of 50% across 5 emotions	20
8.4.2 Emotion detection rate of 1 frame per second.	21
8.4.3 Track user in real-time.	21
8.5 Spring Validation Experiment Evaluation	21
8.6 Strengths and Weaknesses	21
8.6.1 Strengths	21
8.6.2 Weaknesses	22
9. Project Management	22
9.1 Work Breakdown Structure	22
9.2 Schedule	23
9.2.1 Weekly Schedule	23
10. Conclusions	29
10.1 Key lessons learned	29
10.1.1 Training on multiple datasets is vital	29
10.1.2 Data preprocessing is the most time consuming task	29
10.1.3 Robustness of textual predictions	29
10.1.4 Considering using servos instead of steppers	29
10.2 Future Work	29
10.2.1 Early fusion of all three modalities	30
10.2.2 Longer Contextual Information	30
10.2.3 Emotional predictions as a prior for personality predictions	30
10.2.4 Prediction of micro-emotions	30
10.2.5 Using emotional predictions to improve machine responses	30
11. References	31

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 recognize 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 are out of the scope of this project.

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 anthropomorphize 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 externalized 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.P3	Shall Track user	The camera will track the user during the entire time in 2-DOF (Left -Right & Forward -Backward) at a maximum speed of 10 cm/sec along both directions
------	------------------	--

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 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

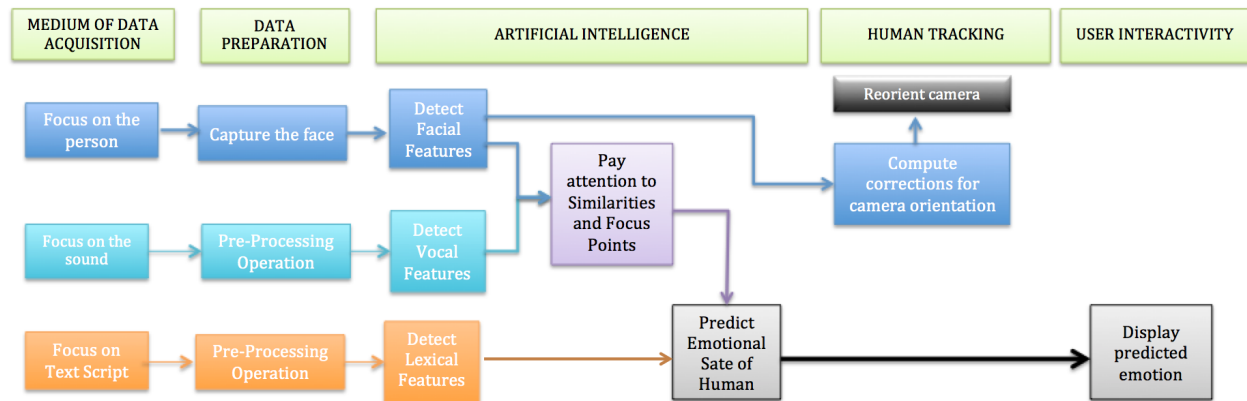


Fig 5.1: Block diagram functional architecture

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 visual data modalities. The learnt common features and focus points (that both modalities show are helpful in predicting emotion) are then fed along with verbal 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. System Level Trade Studies

6.1 Joint Representation versus Coordinated Representation:

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:

Figure 6.1 gives a visual description of joint representation. It has the following characteristics.

- 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 both to represent data and to perform a particular task
- Cannot handle missing data easily although some ways exist to solve this issue

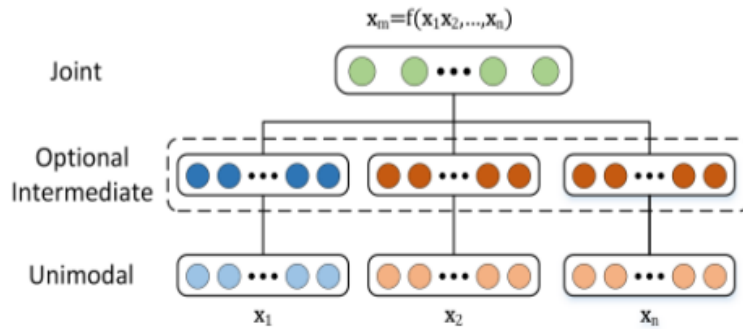


Fig 6.1: Joint Representation[3]

Coordinated Representation:

Figure 6.2 shows the coordinate representation. It has the following characteristics.

- 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

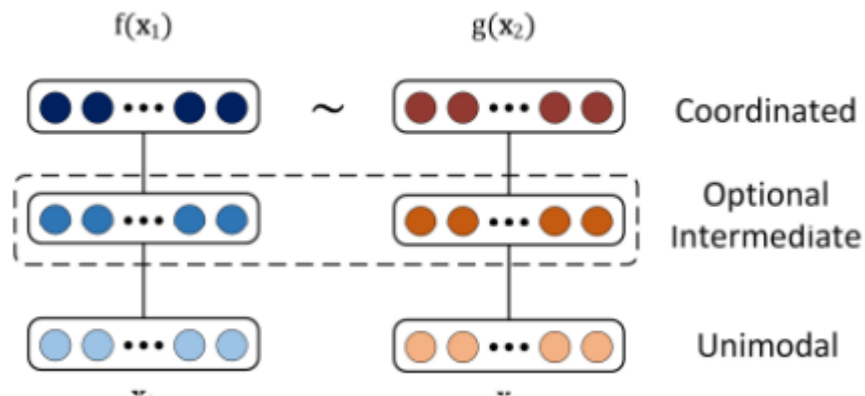


Fig 6.2: Coordinate Representation[3]

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
- Multimodal data at real test time scenarios.

6.2 Early Fusion Versus Late 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.

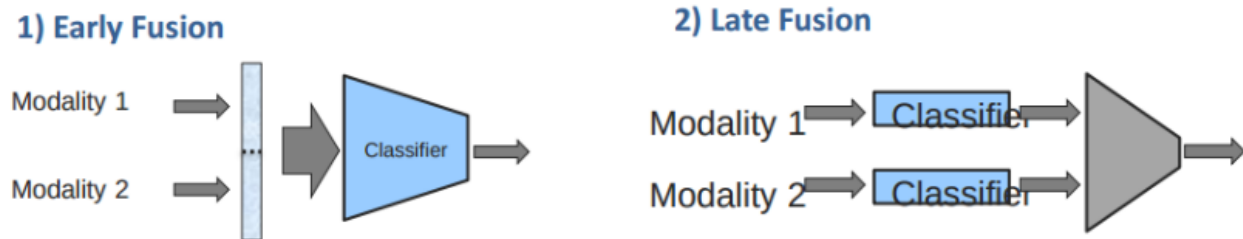


Fig 6.3: Multimodal Fusion Approaches[3]

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.
- If well trained, early fusion type architectures are robust to noise, corrupted or missing modalities.

6.3 Action Unit Features Versus Dense CNN Features:

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. In vision only emotion recognition literature, action units have shown to be directly useful in predicting emotions.

On the other hand, we had the option to use Dense CNN features from Resnet or VGGNet which we trained on emotion recognition. We followed both approaches, on the dataset we got almost similar accuracies and also in the real world. However we were not able to detect the failure points of our vision subsystem since the dense CNN features can't be decoded to what the network is learning. Using the dense CNN features meant a faster training and iteration pipeline due to an end to end deep learning system rather than pre-computation in the action-unit extraction case. However, dense CNN features also had a higher variance among multiple experiments. Owing to lower variance and capability to understand the visual subsystem, we

decided to extract action unit feature intensities rather than dense features which have been detailed in our vision subsystem description.

6.4 Steppers Versus Servos:

The design of the system has been in a state of flux for quite some time. We were originally impressed by the repeatability and precision of stepper motors, but as we came to find out, had to grapple with a host of over issues that would curtail the flawlessness of the integration. Stepper motors have a bit of a pre-actuated lag which makes their response slightly jittery concerning the nature of the input to it, coming in discrete chunks rather than a fluid stream. The steppers were also more prone to overshooting the humans because of this lag, and were also prone to making weird sounding noise which could corrupt the background noise that is being registered by the microphone. Servos on the other hand had the ability to be a lot smoother, without making too much noise, they were also a lot more sturdy than we already thought and were fully capable of moving the Lukabot around. Unlike Steppers, they also did not require a motor driver, and increased the simplicity of our setup.

7. Cyber-Physical Architecture

7.1 Block Diagram

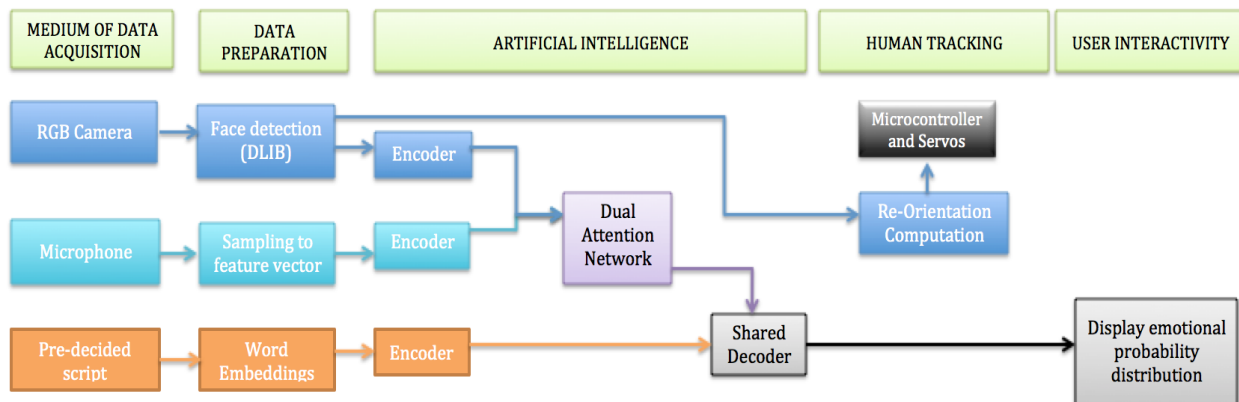


Figure 7.1: Block diagram cyber physical architecture

7.2 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 extract Action Units for each frame along with a bidirectional LSTM along the temporal dimension. For the vocal modality we extract acoustic features every 10 ms and again use a bidirectional LSTM. The extracted features from the verbal and visual 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.

8. System Description and Evaluation

An overview of the system is shown in Fig 8.1.

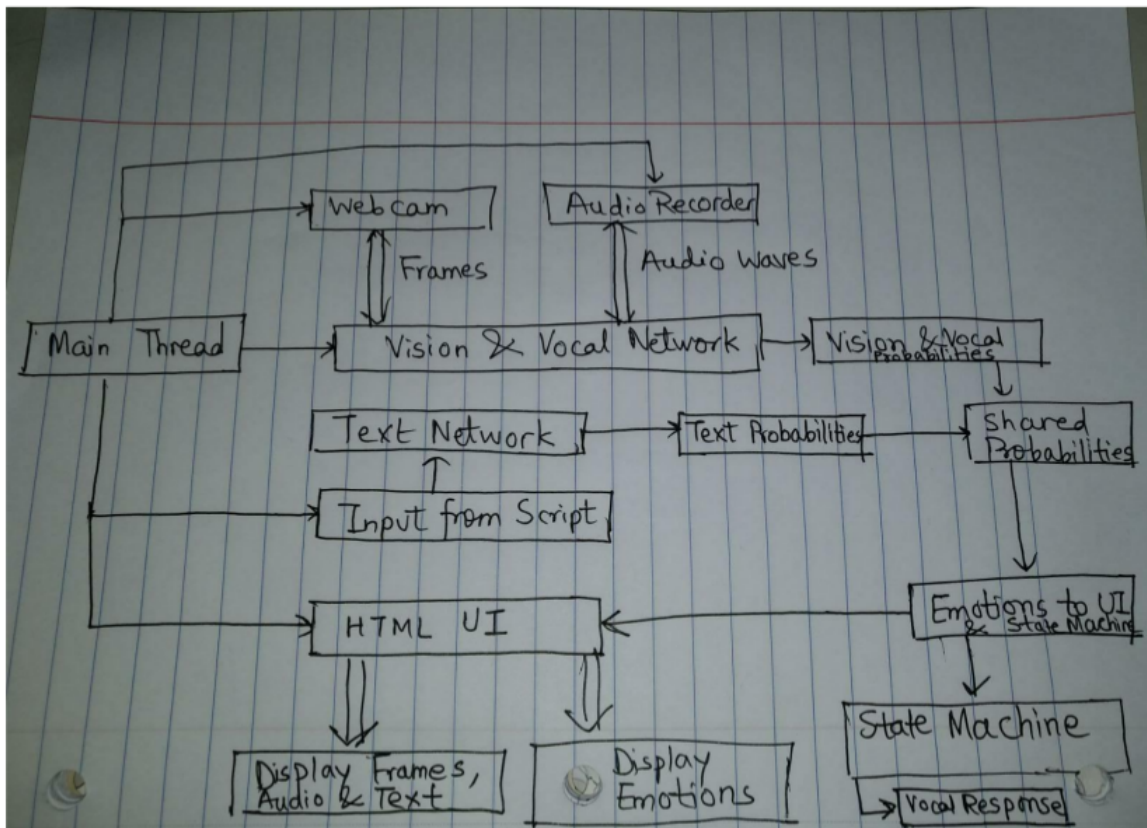


Fig 8.1 : Description of the Integrated System

There is a main thread which launches webcam, audio recorder, the vision + vocal network and the text input. The webcam keeps recording and showing frames to the live GUI. It pushes all the frames to a variable shared across the vision and vocal network. The audio recorder does a similar thing. It records audio and pushes it to a shared array of audio waves. Both these variables are accessed and processed by the vision and vocal network. The text input is taken from a script. This is passed to the text network which pushes its probabilities. The same is the case for the vision + vocal network. These probabilities are accessed by using a shared variable

again in another function. This is again done in parallel since both the encoders calculate at different rates. The shared probabilities function weighs shared probabilities and outputs emotions. The final emotions along with their probabilities, the recorded frames and transcript are sent to the UI. In the coming sections we discuss all of the subsections in detail.

8.2 Subsystem descriptions

8.2.1 Vision Subsystem

The underlying goal of the subsystem is to extract important visual features from raw video feed. Fig 8.2 shows the pipeline for achieving the same.

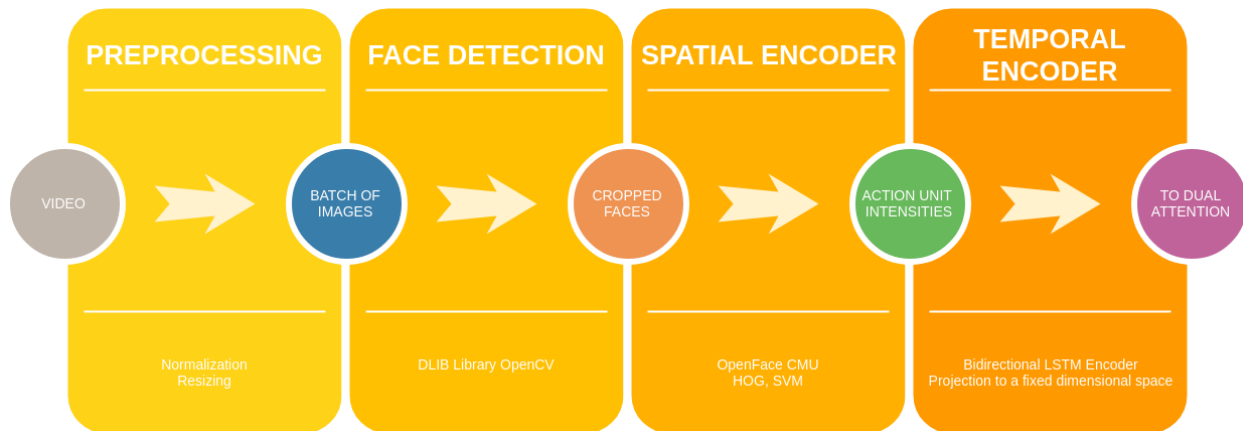


Fig 8.2 : Overview of the Vision Subsystem Pipeline

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. The vision encoder is divided into 2 parts, a spatial encoder and a temporal encoder for learning representations across space and time respectively. Action Unit intensities are then extracted from batches of these images using the spatial encoder. We leverage the CMU OpenFace repository[4] for developing our spatial encoder. In vision-only emotional recognition literature action units have been found to be very critical in recognizing emotions. Action Units(AUs) are fundamental actions of individual muscles or group of muscles. Fig 8.3 touches upon some of the key action units used in emotion recognition.
















<p>AU1</p>  <p>Inner brow raiser</p>	<p>AU2</p>  <p>Outer brow raiser</p>	<p>AU4</p>  <p>Brow Lowerer</p>	<p>AU5</p>  <p>Upper lid raiser</p>	<p>AU6</p>  <p>Cheek raiser</p>
<p>AU7</p>  <p>Lid tighten</p>	<p>AU9</p>  <p>Nose wrinkle</p>	<p>AU12</p>  <p>Lip corner puller</p>	<p>AU15</p>  <p>Lip corner depressor</p>	<p>AU17</p>  <p>Chin raiser</p>
<p>AU23</p>  <p>Lip tighten</p>	<p>AU24</p>  <p>Lip presser</p>	<p>AU25</p>  <p>Lips part</p>	<p>AU26</p>  <p>Jaw drop</p>	<p>AU27</p>  <p>Mouth stretch</p>

Fig 8.3 : Action Units and their visual meaning[5]

The spatial encoder gives us the intensities of all the action units. A snippet of the same is shown in Fig 8.4.

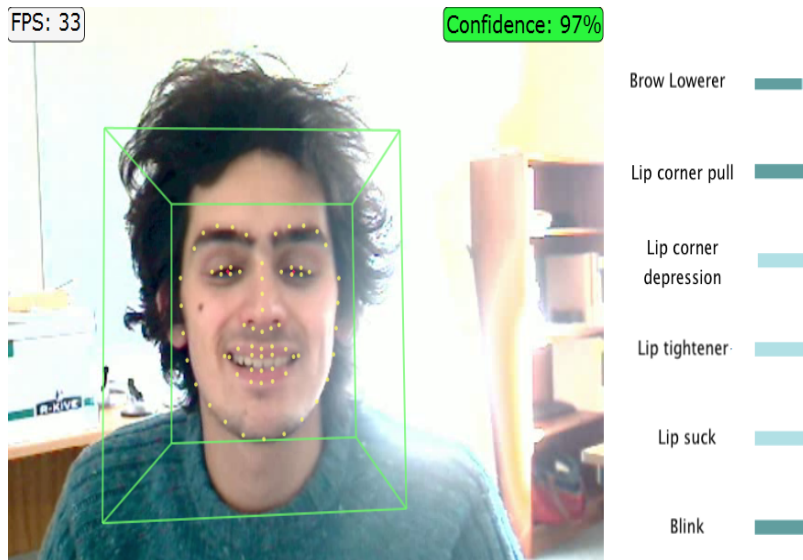


Fig 8.4 : Action Unit Intensity Output of the Spatial Encoder(right) for an example frame(left)[4]

Fig 8.5 draws a relation between the action units and emotion. However we don't use these action units to predict emotion directly but rather feed these features to the temporal Encoder first. The Encoder is a Bi-directional LSTM with 256 hidden layers which extracts temporal information over 3 second sequences. These spatio-temporal features obtained from both the encoders are fed into the dual attention network after combining with the various features from the audio modality described in the next section.

Basic expressions	Involved Action Units
Surprise	AU 1, 2, 5, 15, 16, 20, 26
Fear	AU 1, 2, 4, 5, 15, 20, 26
Disgust	AU 2, 4, 9, 15, 17
Anger	AU 2, 4, 7, 9, 10, 20, 26
Happiness	AU 1, 6, 12, 14
Sadness	AU 1, 4, 15, 23

Fig 8.5 : Action Units and their relation to emotions[6]

8.2.2 Vocal Subsystem

The goal of this subsystem is to extract important vocal features from raw waveform. Fig 8.6 shows a raw audio waveform from which features have to be extracted.

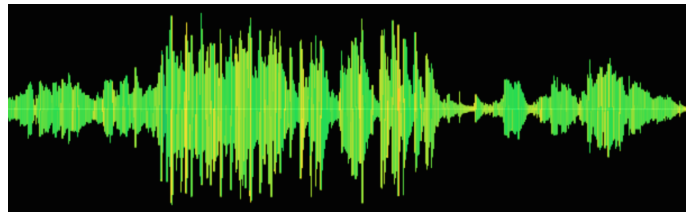


Fig 8.6: Raw Audio Waveform

Fig 8.7 gives an overview of the vocal subsystem pipeline.

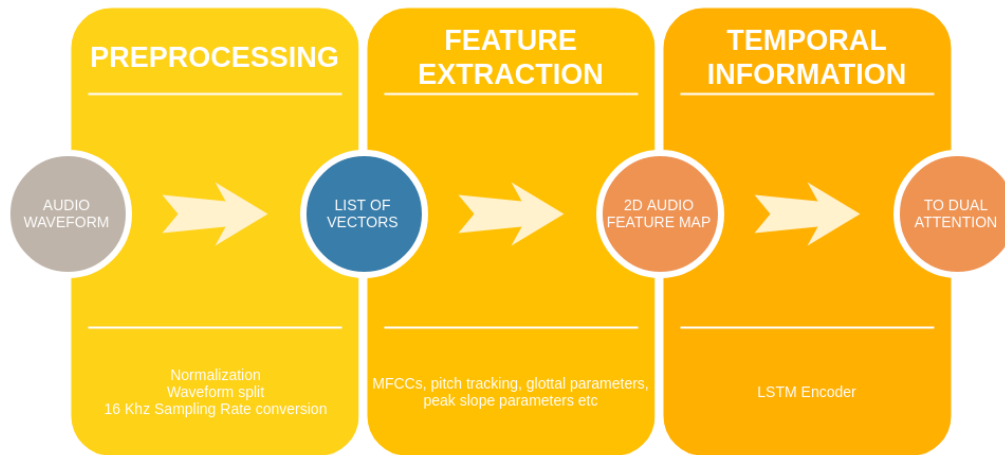


Fig 8.7 : Vocal subsystem pipeline overview

After recording the audio waveform, we resample it to 16Khz and normalize it to get a list of vectors across time. In the feature extraction section for each utterance audio, a set of acoustic features are extracted using COVAREP acoustic analysis framework[7], including 12 MFCCs, pitch tracking and Voiced/UnVoiced segmenting features (using the additive noise robust Summation of Residual Harmonics (SRH) method), glottal source parameters (estimated by

glottal inverse filtering based on GCI synchronous IAIF, peak slope parameters , maxima dispersion quotients (MDQ) , and estimations of the Rd shape parameter of the Liljencrants-Fant (LF) glottal model . These extracted features capture different characteristics of human voice and have been shown to be related to emotions. Fig 8.8 shows the relative importance of top 10 features in predicting emotions. The resultant output is a 2D feature map with rows along time and each column representing a different feature value. A bidirectional LSTM Encoder is then used across time on this 2D Feature map. The representation output from the LSTM is fed into the Dual Attention Network along with the visual modality as discussed earlier.

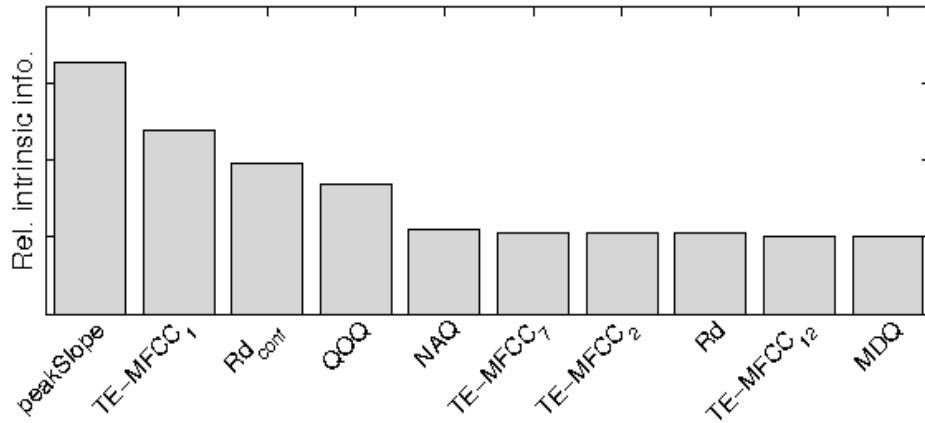


Fig 8.8: Top 10 audio features[7]

8.2.3 Verbal Subsystem

The aim of this subsystem is to extract important learned features from raw text/transcript data. Fig 8.9 shows the pipeline for the verbal subsystem.

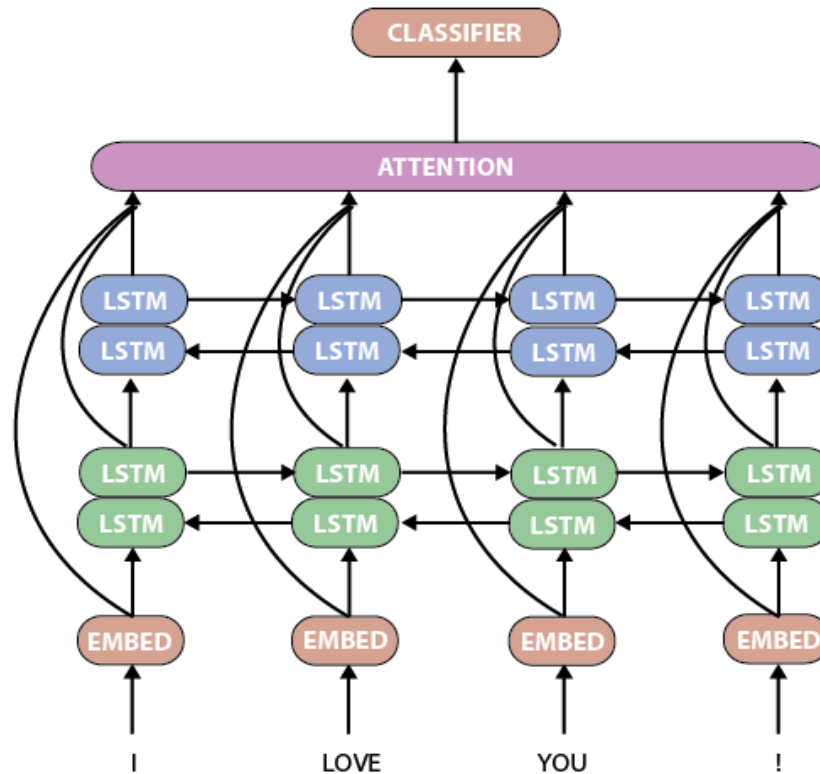


Fig 8.9: Overview of the verbal subsystem[8]

In the preprocessing stage the transcript is divided into words and each word is converted to its word embedding. These are then passed to the bidirectional LSTM with 1024 hidden cells with 512 along each direction. The attention mechanism lets the model decide the importance of each word for the prediction task by weighing them when constructing the representation of the text. For instance, consider the sentence “The weather is so nice today” then a word such as ‘nice’ is likely to be very informative of the emotional meaning of a text and it should thus be treated accordingly. The attention model should give less weightage to words like “today” or “weather”. On a low level, the attention model for the text modality is a simple weight matrix over all the words. The importance of each word in a sample is learnt during the training phase. Lastly, the representation vector for the text is found by a weighted summation over all the time steps using the attention importance scores as weights. This representation vector obtained from the attention layer is used as input to the final softmax layer for classification. As is shown in the above figure there are also skip connections between the Bi-LSTM and the attention as well as between the embeddings and the attention. The original authors of the paper argue that these skip connections were useful for transfer learning on the text modality.

8.2.4 Attention Model

The goal of the attention model is to learn similarities between the vocal and the visual 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. For our project

we employ a dual attention network. Dual Attention Networks (DANs) which jointly leverage visual and textual attention mechanisms to capture fine-grained interplay between vision and audio modalities. DANs attend to specific regions in images and waveforms through multiple steps and gather essential information from both modalities. Based on this framework, we work with r-DAN or reasoning DAN. The reasoning model allows visual and audio attentions to steer each other during collaborative inference, which has been shown to be useful for tasks such as Visual Question Answering (VQA). We have found that r-DANs are also helpful in emotion recognition task giving us a higher performance over a single human rater on the CREMA-D dataset. Fig 8.10 shows a high level overview of the attention model.

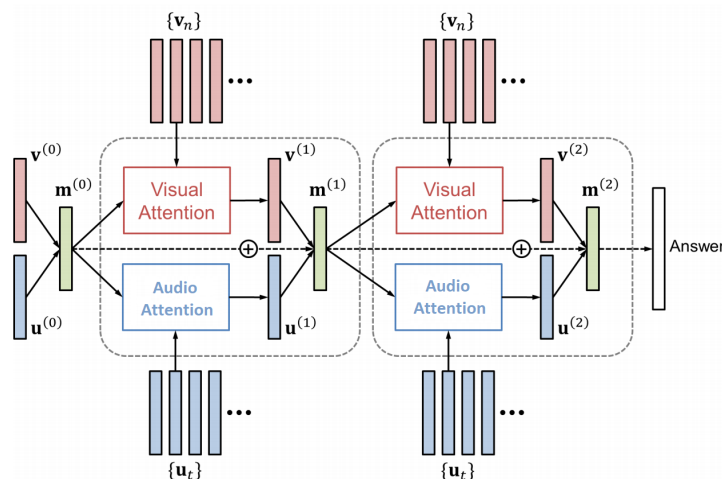


Fig 8.10:Reasoning Dual Attention Network[9]

On a low level the attention model, shown in fig 8.10, 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 dual attention is connected with a fully connected layer to give the final audio-visual predictions. These softmax probabilities are later combined with the text probabilities to give the final emotional output.

8.2.5 UI

The goal of this subsystem is to show the performance of the system while also showing the corresponding video recording and the text transcript spoken. Fig 8.11 shows a snapshot of our user interface subsystem.



This is one of the most impressive talks that I have been to.

Actual Emotion : Happiness

Results



Prominent Emotion : Happiness

Fig 8.11: Results of the integrated network

The UI outputs probabilities of each emotion for the audio-visual and the text network as well as shown in Fig 8.12

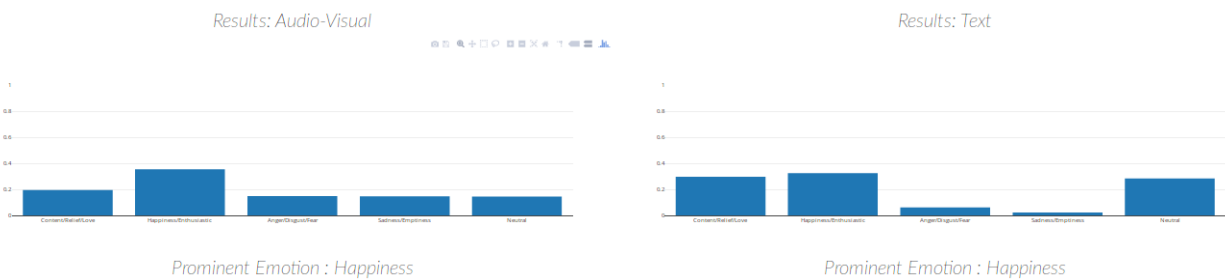


Fig 8.12: Results of the audio-visual and text predictions

8.2.6 Robot Hardware

The goal of this subsystem is to track the face of a human as they move about within the field of view of the camera of the robot. Interactions rarely feature completely static human motion, and so it is of the utmost importance that the visual data pipeline hardware (the camera) is able to pan from side to side as the human walks and leans from side to side, and also tilt upwards and downwards as the human stands up, sits down, leans in, recedes away from the robot. A static camera would lose such a person in a truly natural interaction. This system works by detecting the center of a face and relaying the pixel coordinates to the computer so that instructions can be sent to Servos so that they rotate until the face is centered in the image. An Arduino is the intermediate micro-controlling agent that acts between the face coordinate detection of the PC, and the actuation of the Servos. This a small robot with a height of 32cm from base of the stepper to the top of the webcam. The assembled robot is shown in Fig 8.13

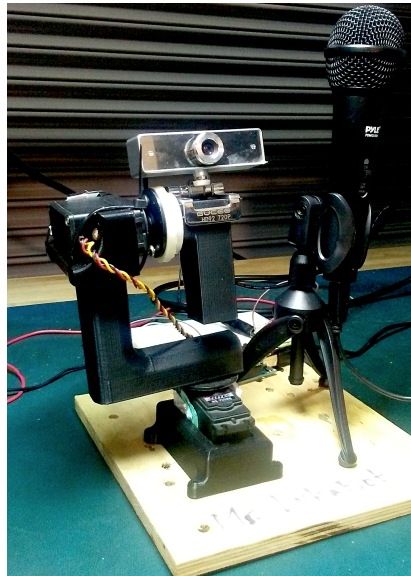


Fig 8.13: Assembled LukaBot

8.3 Modeling, Analysis and Testing

8.3.1 CAD Modeling of Face Tracker

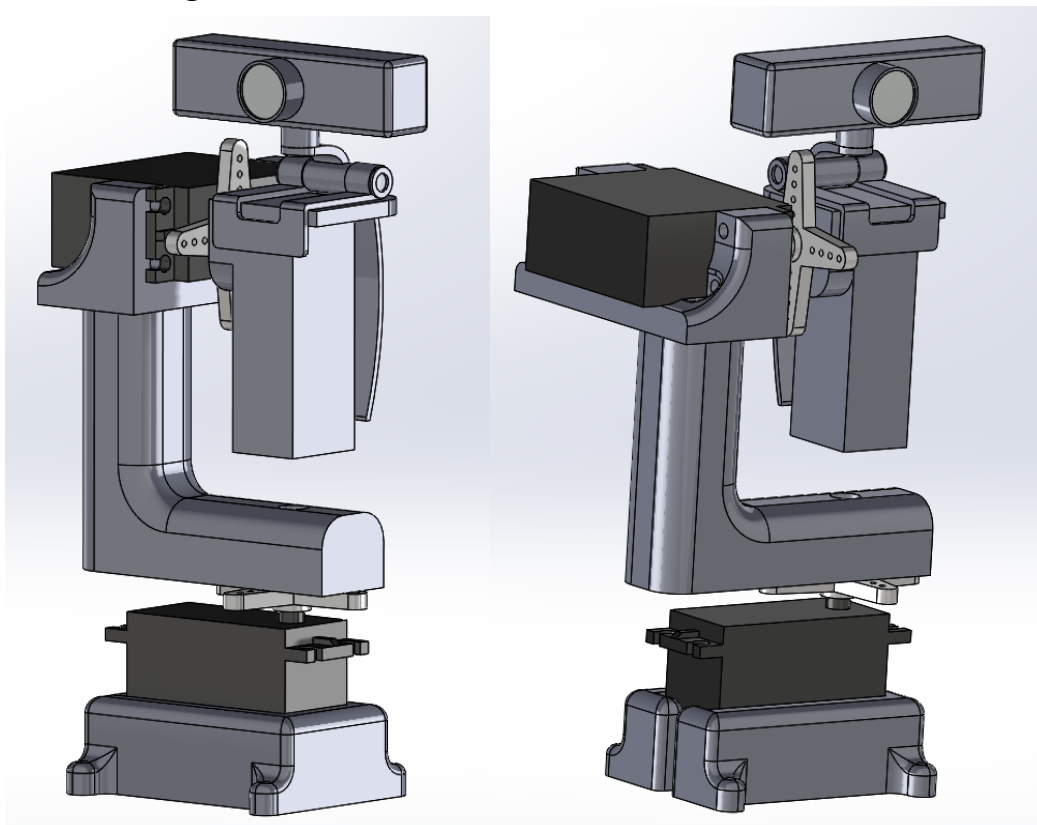


Figure 8.14 shows the CAD Model of the 2-DOF LukaBot.

8.3.2 Experimental Results

Table 8.1 benchmarks our average results on real tests conducted before, during and after the SVE and SVE encore. The reported results were measured using 3 human raters (our team) as ground truth across 5 emotions. These 5 emotions are content, happiness, sadness, anger and neutral.

Table 8.1: Results on real tests

Modality	Accuracy(across 5 emotions)
Audio + Vision	55%
Text	59%
Audio+Vision+Text	71%

Table 8.2 shows our results on the CREMA-D dataset against 6 emotions. These emotions are happiness, anger, fear, disgust, sadness and neutrality. The model with the highest accuracy (at the bottom of the table) was chosen as our final model. We can see that OpenFace features also slightly outperform dense CNN features as argued in our trade studies section. We also beat the

human-level performance on this dataset. We also show the jump in performance upon combining modalities clearly.

Table 8.2: Results on CREMA-D datasets across 6 emotions

Model	Mode	Acc.
Human performance	Audio	40.9%
COVAREP Features + LSTM Decoder	Audio	41.5%
OpenFace Features + LSTM Decoder	Vision	52.5%
Resnet-18 + LSTM Decoder	Vision	54.8%
Resnet-18 + (COVAREP Features + LSTM) +Gated Attention	V+A	58.0%
Human performance	Vision	58.2%
Human performance	V+A	63.6%
Resnet-18 + (COVAREP Features + LSTM) +Dual Attention	V+A	63.6%
(OpenFace features + LSTM) + (COVAREP Features + LSTM) +Dual Attention	V+A	65.0%

We use the best audio, vision and audio-vision models to run more experiments on a different dataset. The RAVDESS dataset has 8 different emotions (neutral, content, happy, sad, angry, fearful, disgust, surprised) and two different types of recordings, one being normal speech and other being singing. The results for both the types are shown in table 8.3 and 8.4 respectively.

Table 8.3: Results on RAVDESS dataset for normal speech(across 8 emotions)

Model	Mode	Acc.
COVAREP Features + LSTM Decoder	Audio	41.25%
OpenFace Features + LSTM Decoder	Vision	52.08%
(OpenFace features + LSTM) + (COVAREP Features + LSTM) +Dual Attention	V+A	58.33%

For the singing recording there are only 6 emotions (neutral, content, happy, sad, angry, fearful). Although the model trained on singing recordings are not used, it was useful to see the transfer of learning from normal speech to singing. It is evident in the results that since singing audio differs a lot from normal speech audio our accuracies are almost close to random. The Dual modality accuracies also fall down to single modality accuracy levels. We conclude that the dual attention

is very robust to corruption of modalities (speech to singing) and would behave almost like a single modality despite it being an early fusion type of architecture. This has been one of the reasons for choosing early fusion in our trade studies before.

Table 8.4: Results on RAVDESS dataset for singing (across 6 emotions)

Model	Mode	Acc.
COVAREP Features + LSTM Decoder(pretrained on speech recordings only)	Audio	21.59 %
OpenFace Features + LSTM Decoder(pretrained on speech recordings only)	Vision	50.56 %
(OpenFace features + LSTM) + (COVAREP Features + LSTM) +Dual Attention(pretrained on speech recordings only)	V+A	51.13 %
(OpenFace features + LSTM) + (COVAREP Features + LSTM) +Dual Attention(with fine tuning on song recordings)	V+A	67.61 %

8.4 SVE Requirements

8.4.1 Emotion detection accuracy of 50% across 5 emotions

All the various real-life emotions were compartmentalized into 5 distinct emotion neighborhoods/buckets for identification as shown in the above figure.

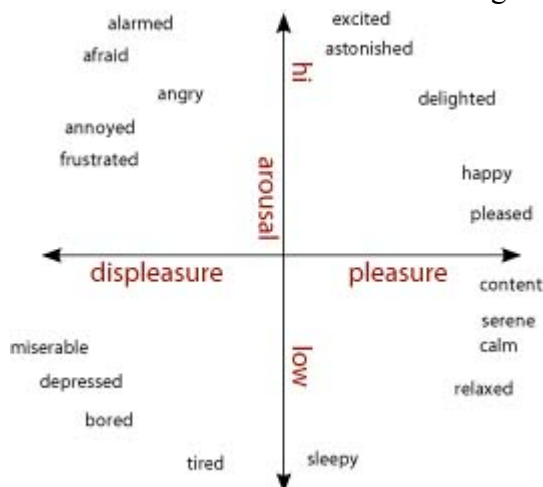


Fig 8.14: Valence-Arousal Emotion Chart[10]

The first bucket is happy/elated/excited shown in the top right quadrant. The second neighbourhood is sad/depressed/gloomy given by the bottom left quadrant. The third cluster is hate/alarmed/fear/anger in the top left quadrant. The fourth one is calm/content/relaxed in the bottom right quadrant and the last one is neutral in the center. The requirement was that the neural network should perform at an average accuracy of 50% on real life tests.

8.4.2 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.

8.4.3 Track user in real-time.

While enacting emotions, the user is free to maneuver around the room at a pace of 10 cm/sec both forward and backward as well as left and right. Lukabot should be able to track the user throughout the demo. Successful tracking is defined as the ability of the bot to keep the human inside the frame at all times.

8.5 Spring Validation Experiment Evaluation

The goal of experiment was to test the performance of Lukabot on recognizing emotions in real-time. The user is free to move from left to right at 10 cm/sec and forward and backward at 10 cm/sec. The success criterion included a processing performance of 1 frame per second and an average accuracy of 50% across 5 emotions. These 5 emotions were content, happy, anger, sadness and neutral. For the SVE we achieved an accuracy of 59% and on the day of the SVE Encore we achieved an accuracy of 71%. The jump in the accuracy is partly attributed to training on another dataset as well as no audio responses from the Lukabot interfering with the actor's audio adhering to feedback received during the SVE. The system was also able to track the user at all times.

8.6 Strengths and Weaknesses

8.6.1 Strengths

- **Processing Performance:**
The speed of the network at which it predicts is around 30 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.
- **Tracking Performance:**
We are able to achieve successful tracking 100% of the time at speeds of 10 cm/sec left to right and 10 cm/sec forward and backward. We also feel that the tracking is very smooth and exactly what we envisioned while designing this subsystem.
- **Overall Emotion Prediction:**
We were able to achieve reasonable overall prediction for the emotion recognition subtask. We are able to achieve real world accuracies of around 60-75% depending on the nature of acting by the human. The system is also able to recognize all emotions with high confidence when they are clearly not subtle.
- **Satisfactory confusion matrix:**
Whenever the integrated system is wrong, it mostly confuses between content and happy or sadness and neutral etc. It very rarely confuses between two emotions of opposite

connotations like happy and anger or content and sadness. To quantify this observation, we use another metric called the top-n accuracy. This implies that we consider the classification as correct if the probability of the correct class lies among the top-n sorted probabilities. All our earlier results report the top-1 accuracy. Therefore, although our accuracy is around 70%, our top-2 accuracy hovers around in the 85-90% range.

8.6.2 Weaknesses

- Separation of audio-vision and text modalities:
The high-level audio and visual modality features were fused using the dual attention network. Through our experiments and ablation studies we have seen that by doing this we achieve more robustness even when one of the modalities is corrupted to a certain extent (like blurry face, mismatched audio recording etc). This is because the Dual Attention is trained to attend to important features across the audio-visual space. Therefore our accuracy levels fall back to single modality levels in such scenarios. However for integrating the audio-vision and text modalities we use a simple averaging of softmax predictions. When the audio-vision or the text modality is corrupted this sometimes can give a lower accuracy than individual modalities combined. This can be fixed by using a triple attention or other fusion methods across all three modalities. However in the interest of time, this avenue was not pursued.
- Short Contextual Information:
Our current system predicts emotions for every 3 second chunk of audio-visual and text modality. For each segment, it does a totally new prediction and doesn't take into account the emotions in the previous segments. It only considers the time - information within those 3 seconds to give a prediction. This worked fine for the demo where the system must be demonstrated for a short period of time, but in real life emotions are more invariant to change every 3 seconds. Therefore a system which takes longer contextual information into consideration would perhaps be more useful.

9. Project Management

9.1 Work Breakdown Structure

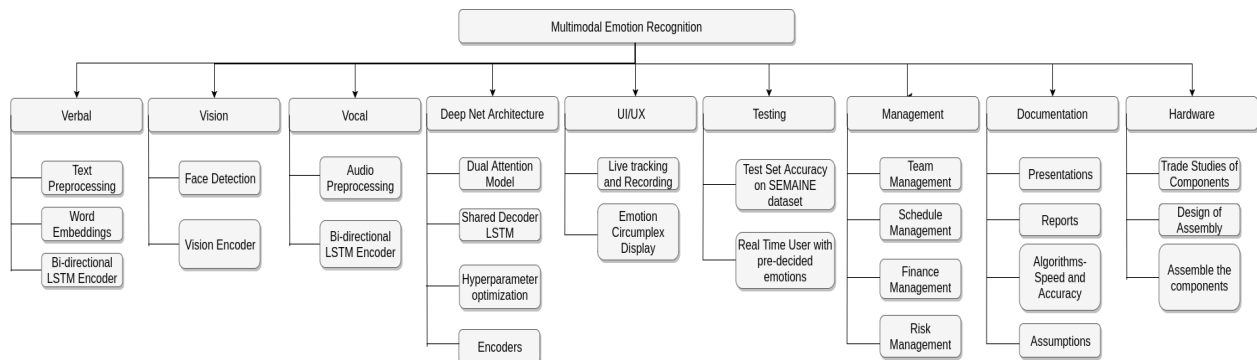


Fig 9.1: Three-level Work Breakdown Structure

As shown in the figure 9.1, 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.

Figure 9.2 (below) shows the work pages in a more detailed manner annotated by semesters they're scheduled to be finished in. Here is how you read this figure: At the bottom right, there is a key that explains the meaning of the colors red, green and yellow. These colors appear everywhere else in the figure as thin left sided borders to each cell. The actual color code of each work package (purple, orange, blue, pink, gray etc) that fills up the cells is completely irrelevant and is only there to give each work package its own separate color. Within each work package, there are darker or lighter shaded cells, the darker ones are only heading cells, with uncolored ones being second order work breakdown of those heading cells.

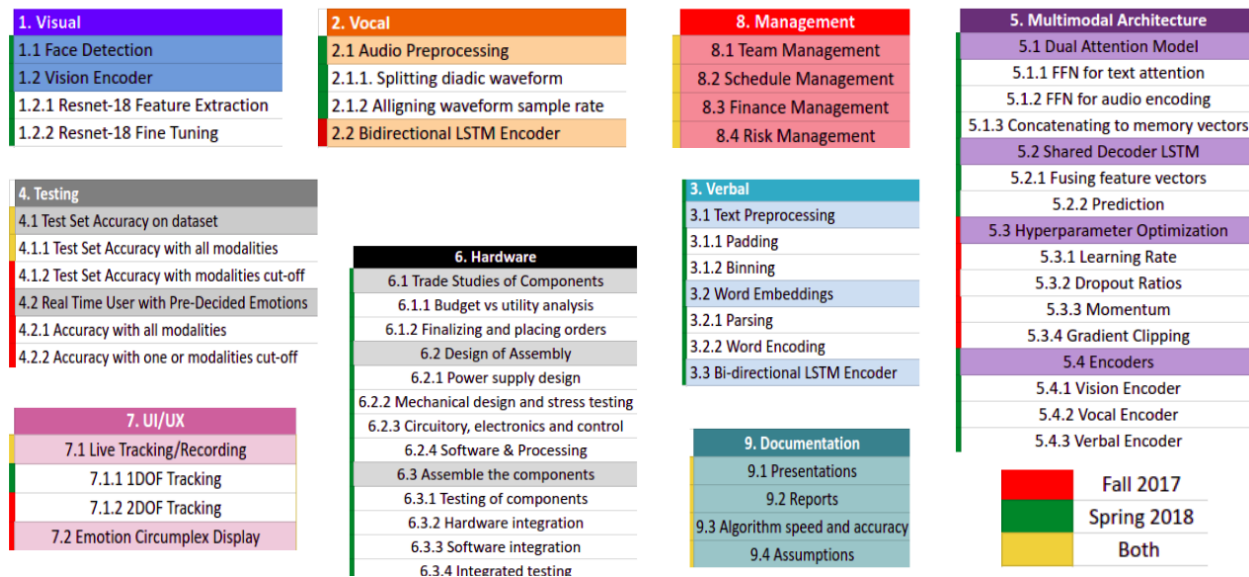


Fig 9.2: Work packages

9.2 Schedule

9.2.1 Weekly Schedule

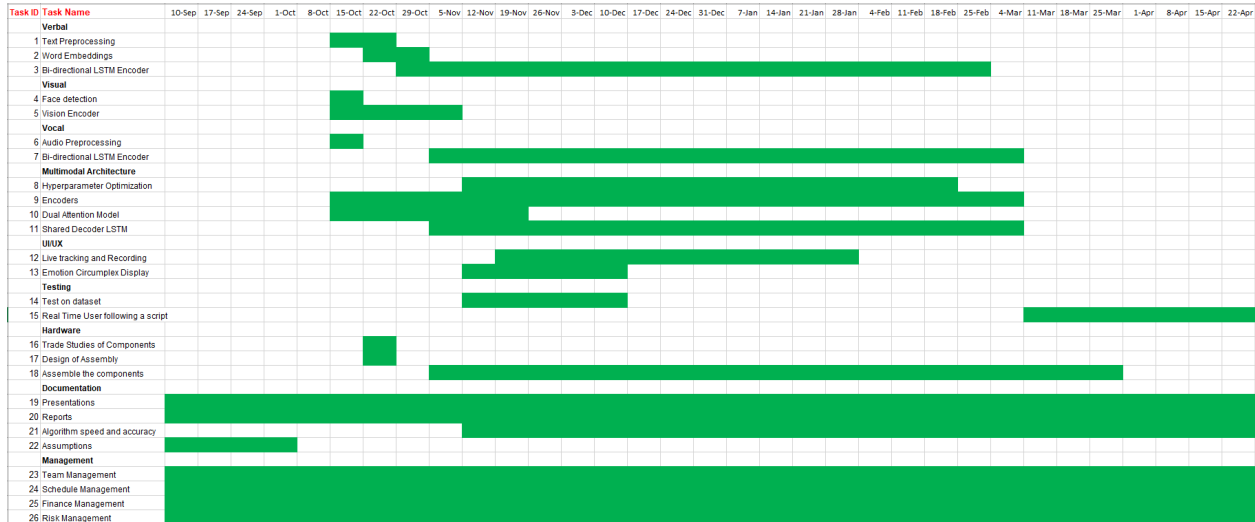


Fig 9.3 Gantt Chart

Weekly schedule of activities have been documented and were tracked in Fig 9.3.

9.3.2 Spring Validation Experiment

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 15-30 cm in front of the camera.
- Single active subject, face may move at pace of
 - 10 cm/sec (Left - Right)
 - 10 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 during the test at all times.

9.4 Budget

9.4.1 Refined Parts List

Table 9.1: 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

9.4.2 Budget Summary

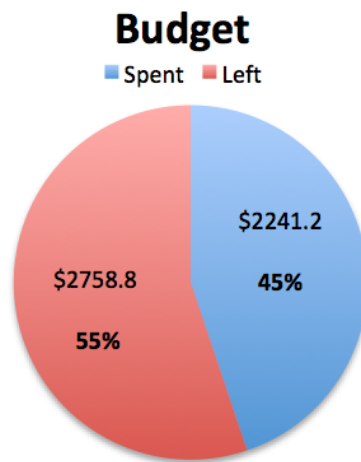


Fig 9.4 Budget status

At the start of the project, we were given \$5000 to spend. Due to the extreme lack of actual hardware in this project, we were a lot more adventurous with the items we obtained, and more inclined on purchasing items to empirically validate them as opposed to waiting too long doing theoretical validations. We have purchased items that were not directly associated with the actual Lukabot such as additional hard drives for the workhorse computer in the MRSD lab so that we could store the great big quantity of data that we would need to train our algorithms on. We also purchased a large quantity of contingency parts in case some broke down, and so ordered things like Steppers, Servos, Arduinos in a large Bulk which of course largely contributed to the final budget. We also order items that we ended up not using such as Microsoft Kinect Cameras (with contingent ones) as well as oversized Steppers.

9.5 Risk management

9.5.1 Risk Identification

We have identified the following major risks and have tracked them closely:

1. Lack of diversity in datasets

2. Delay due to high training time
3. Hardware breakdown
4. Low accuracy

We have then modelled the impact-likelihood ratings of these risks as follows:

		Impact →				
		Negligible	Minor	Moderate	Significant	Severe
Likelihood ↑	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 9.5: Risk impact vs likelihood table for major risks

9.5.2 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	
<ol style="list-style-type: none"> 1. Exposing the network to varied datasets 2. Using networks pretrained on large datasets like ImageNet 3. Introducing invariance(eg using landmarks over pixels) 			<ol style="list-style-type: none"> 1. Higher time requirement 2. Importing pre-trained models and weights 3. Tweaking network architecture 	

Fig 9.6: 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	
<ol style="list-style-type: none"> 1. Prototyping on smaller subset of training data 2. Parallelizing tasks: parallel development, parallel training 3. Optimise code for time and space complexity 			<ol style="list-style-type: none"> 1. Shorter development cycles 2. Pooling and threading, running on multiple cores 3. Better running time 	

Fig 9.7: 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	
<ol style="list-style-type: none"> 1. Extensive testing and debugging after integration 2. Stock lots of spares for contingency 3. Make risk-prone hardware modular to enable easy replacement 			<ol style="list-style-type: none"> 1. Higher cost 2. Compact amenable packaging and integration 	

Fig 9.8 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	
<ol style="list-style-type: none"> 1. Downsize requirements 2. Aggressive, intuition and evidence driven experimentation 3. Extensive literature review 			<ol style="list-style-type: none"> 1. More time for training and tuning 2. More time for literature review 3. Robust networks 	

Fig 9.9 Risk card for Risk ID 4

Risk mitigation cards for the four major identified risks have been given in figures 9.6, 9.7, 9.8 and 9.9.

10. Conclusions

10.1 Key lessons learned

10.1.1 Training on multiple datasets is vital

Since the final product has to perform on real world conditions it was 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. Our encoders were pre-trained on 5 datasets and the attention model along with the end to end network were trained on 2 datasets. We found that our encoders were pretty reliable and the attention model worked very well as detailed in the experiments section.

10.1.2 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.

10.1.3 Robustness of textual predictions

Textual predictions are not susceptible to acting, quality of recordings, lighting conditions, facial orientation or other biases not observed in the dataset. This makes the text modality robust to variance among experiments with the same transcript. It doesn't change its performance by a large factor from dataset to real world conditions as compared to other modalities.

10.1.4 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 was be less dependent on batteries working well, meaning that there were less connections that run the risk of shorting.

10.2 Future Work

10.2.1 Early fusion of all three modalities

The audio and visual modalities were fused early using the dual attention model. However, the text modality was fused at the end with the audio-visual modality. Using a triple attention to fuse all the three modalities together would lead to more robustness in predictions when one or more modalities are corrupted.

10.2.2 Longer Contextual Information

Currently, we only take the last 3 seconds of temporal information to predict emotions. In the future, predictions should take in longer time intervals into consideration as well. A moving average would simply smoothen out the predictions but rather a model which learns the causal nature of the time space is important to be learnt.

10.2.3 Emotional predictions as a prior for personality predictions

Emotional responses of a person over a period of time can be used to predict the personality of people and vice versa too. For instance, optimistic people are happier on an average than pessimistic humans. Some people emote subtly while others not so much. Personality and emotions go hand in hand. Therefore there should be a joint model which learns the personality of the person as well as his emotions together. This is commonly referred to as multi-task learning and has been shown to be very successful when the tasks are similar. The idea behind multi-task learning is that it usually gives better predictions on both the tasks individually rather than having separate models for each task given that the tasks are similar.

10.2.4 Prediction of micro-emotions

There are many more emotions other than the 5 major ones which we have identified in our project. Due to lack of datasets which have emotions labelled with a high granularity ($\geq 10-15$), prediction of micro-emotions was out of the scope. But in the future, if such a dataset can be made or acquired then the network would predict emotions on a finer level. On other machine learning tasks, it's been shown that increasing the granularity leads to better predictions even on the same dataset. For instance, if we had only 3 classes; positive, negative and neutral instead of 5 classes; happiness, content, sad, anger and neutral, then our predictions would be worse off for the former 3 class problem than if we lumped the 5 classes *after* the predictions into positive(happy and content), negative(sad and anger) and neutral. The only difference is that in the former case, the network directly predicts 3 classes while in the latter case, we predict 5 classes but later lump them manually to compare against the 3 class predictions. This has been shown to be true in many areas of deep learning and therefore is a viable future alternative.

10.2.5 Using emotional predictions to improve machine responses

Emotional intelligence guides a lot of human-human interaction, but is completely absent from the human-robot/ human-computer interaction arena. Ultimately, in order to give a human

flair to machine responses we need to solve the ability to not only perceive nuances in language but also perceive the subtle emotional cues from facial expressions, tone of voice, and sentimental semantic meaning of the words we use. The perception of these cues is what we humans use when conversing with one another in order to build rapport or build comfort. Our own emotions as well as the reading of the emotions of others guides the response avenues we go down, and using this with machines will help avoid undignified social gaffes on the part of the machine such as giving suggestions or advice that clashes with the mood of the conversation. Just having a chatbot that can hold a conversation may not provide that cathartic experience that comes about from a good talk. In order to vent, or have a conversation that uplifts your spirit or makes you melancholic, an agent needs to monitor how the nature of the non-verbal interaction changes as the verbal conversation happens.

11. 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] <https://arxiv.org/pdf/1705.09406.pdf>
- [4] <https://github.com/TadasBaltrusaitis/OpenFace>
- [5] <https://www.ecse.rpi.edu/~cvrl/tongy/aurecognition.html>
- [6] Ghayoumi, Mehdi & Bansal, Arvind. (2016). Unifying Geometric Features and Facial Action Units for Improved Performance of Facial Expression Analysis.
- [7] G. Degottex, J. Kane, T. Drugman, T. Raitio and S. Scherer, "COVAREP — A collaborative voice analysis repository for speech technologies," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, pp. 960-964.
- [8] <https://arxiv.org/pdf/1708.00524.pdf>
- [9] Nam H, Ha JW, Kim j, "Dual Attention Networks for Multimodal Reasoning and Matching", arXiv:1611.00471, <https://arxiv.org/abs/1611.00471>
- [10] https://en.wikipedia.org/wiki/Emotional_granularity