PROJECT REPORT

Cover

+ Name of the project
+ Project Partners
    - Manas Kale :: 403037
    - Rohit Patankar :: 403053
    - Shubham Punekar :: 403061
    - Saurabh Shirodkar :: 403074
+ Company Name

    IBM
+ Name of the internal guide and co-guide
+ Name of the external guide

Certificate of completion (with MIT Logo)

Abstract

In this project, we have focussed on inculcating machine emotional intelligence in order to improve human computer interaction for the elderly. Emotion detection is a challenging problem partly because the it is difficult to determine the features that might be relevant for the task, and partly because emotive states are ovelapping and not mutually exclusive. Emotional state of a person at any given time is not categorical, but a value from a wide spectrum of emotions. We have followed the available emotion scales (Ekman, Plutchik) and determine the most common subset of emotions which can generically capture the emotive state of a person. We focus on an ensemble approach of utilizing voice, spoken content as well as the facial features to detect emotional state of a person. This emotional state of a person is used in reinforcement loop to improve interaction with NAO robot. We propose to utilize a combination of deep neural network and extreme learning machine for tone analysis, convolutional neural network for facial feature analysis and multinomial naive bayesian classification for spoken content analysis. Finally, the detected emotional state is used along with a neural responding machine to generate a response on the NAO robot.

Keywords : Human Computer Interaction, Machine Emotional Intelligence, Image Processing, Natural Language Processing, Speech and Audio Processing, Machine Learning, Affective Computing.

Acknowledgement

Contents

List of Tables

List of Figures

1 Problem Statement

Improve Human Computer Interaction with Machine Emotional Intelligence using Nao Robot

- to recognize subjects based on their facial features and voice,
- to analyse the facial features, speech text and the tone of speech to detect emotions,
- to generate and emotive score based on a weighted scores from former analyses,
- to generate a context appropriate response on the robot.

2 Project literature survey

1. *Jeong-Sik Park, Gil-Jin Jang* - **Implementation of Voice emotion Recognition for Interaction with Mobile Agent , ACM 2014** [1]
   The paper proposes a simple smartphone interface framework which consists of detection of human voice, extraction of emotional features and identification of an emotional state. Energy based approach for detection of human voice is selected in which if continues estimates of spectral energy of consecutive frames exceeds a pre-determined threshold, the region is regarded as starting point of voice signal. The pitch, log-energy, and "Mel-Frequency Cepstral Coefficient (MFCC)" are selected for extraction of emotional features which make a feature vector sequence. Acoustic features vectors extracted are analyzed and compared with patterns for each emotion type. Guassian Mixture Model is the classification algorithm used. This approach achieved 70.1% correctness within 1s response time. The future work suggested is applying proposed concepts for human-machine interaction in personal agent applications.

2. *Yu Gu, Eric Postma, Hai-Xing Lin* - **Vocal Emotion Recognition using Log-Gabor Filters , ACM 2015** [2]
   The propsed work utilizes 2d Gabor filters in order to decompose the associated spectogram in order to perform a spectro-temporal analysis of affective vocalizations. Instead of including all potentially relevant features which leads of dimensionality problem and subsequent degradation of performance, the work uses "feature learning" in which relevant features are automatically obtained from raw speech signals. However this leads to considerable computational resouces. Hence, no. of features is kept to minimum. By performing analysis on local specto-temporal structure, the spectrogram is treated as an image and standard image processing is implemented. Comparative evaluation of MFCC and LPCC features, untuned and tuned Gabor filters and all above combinations is done. SVM is used as a classifier. The confusion matrix for performance using tuned Gabor Filters provide a maximum of 91.6% accuracy whereas a combination of acoustic features and Gabor filter provides 93.5% accuracy.

3. *Gloria Zen, Elisa Ricci, Nicu Sebe* - **Unsupervised Domain Adaption for Personalized Facial Emotion Recognition, ACM 2014** [3]

A personalization approach is proposed in which only unlabeled target-specific data are required. A new method to represent the source sample distribution based on only Support Vectors of source classifiers is proposed. Regression framework is used to learn a mapping between a marginal distribution of the data points associated to a given person and the parameters of his/her personalized classifier which is represented by a set of Support Vectors of linear classifier in the source case and by all unlabeled data points in the target case.

4. *Ahmed Mustafa Mahmoud, Wan Haslina Hassan* - **Determinism in Speech Pitch Relation to Emotions, ICIS 2009** [4]
   A deterministic rule-based text-to-speech emotional synthesis approach is proposed to generate emotional speech using semitonic interval-driven rules. Emotional speech samples are analyzed and intervals are extracted using praat tool. Objective evaluation compares sysnthesized voice to natural voice and calculates difference as an error function by considering mean square error as a measure of similarity. New emotional states may be defined using same proposed approach. Algorithms that integrate two or more emotional states may be combined to generate a variety of complex emotions.

5. *Nancy Semwal, Abhijeet Kumar, Sakthivel Narayanan* - **Automatic Speech Emotion Detection using Multi-Domain Acoustic Feature Selection and Classification, IEEE 2015** [5]
   The proposed approach concentrated on determining emotions from speech signals. Various acoustic features such as energy, zero-crossing rate(ZCR), fundamental frequency, Mel Frequency Cepstral Coefficient are extractedfor short term, overlapping frames derived from the input signal. A feature vector for every utterance is then constructed by analyzing mean, median, etc. over all frames. Sequential Backward Selection is used with K-fold cross validation to select a subset of useful features. Detection of emotions is done by classifying respective features from the full candidate feature vectors into classes, using either a pre-trained SVM or a Linear Discriminant Analysis classifier. Accuracy of 80% was obtained when tested on EmoDB dataset.

6. *Lei Pang, Chong-Wah Ngo* - **Multimodal Learning with Deep Boltzmann Machine for Emotion Prediction, ACM 2015** [6]
   In contrast to existing works which concentrate on either Audio, text or video, a joint density model is proposed over the space of multi-modal inputswith Deep Boltzmann Machine. The model is trained directly

on user-generated Web videos without any labelling effort. Multiple layers of hidden units and multiple modalities make learning difficult, hence learning is split into 2 stages. First, each RBM component is pre trained using greedy layerwise strategy. Then, learnt parameters are used to initialize the parameters of all layers in DBM and then the multimodal DBM is trained to finetune different modalities in a unified way. A major factor is that the deep architecture enlightens the possibility of discovering highly non-linear relationships between low-level features across different modalities. A performance improvement of 7.7% in classification accuracy is observed.

7. *Benjamin Guthier, Rajwa Alharthi, Rana Abaalkhail, Abdulmotaleb El Saddik* **Detection and Visualization of Emotions in an Affect-Aware City, ACM** [7]
   In the proposed work, emotions are represented as four-dimensional vectors of pleasantness, arousal, dominance and unpredictability. In the training phase, emotion word hashtags in the messages are used as the ground-truth emotion contained in a message. A neural network is trained by using the presence of words, hashtags and emoticons in the message as features. During the live phase, these features are extracted from geo-tagged Twitter messages and given as input to neural-network. The detected emotions are aggregated over space and time and visualized on a map of the city.

8. *Huaizu Jiang, Erik Learned-Miller* - **Face Detection with Faster R-CNN** [8]
   Most approaches to face detection are still based on the R-CNN framework , leading to limited accuracy and processing speed. In this paper, investigations regarding the application of Faster R- CNN which has demonstrated impressive results on various object detection benchmarks, to face detection have been made. By training a Faster R-CNN model on the large scale WIDER face dataset, state-of-the-art results on the WIDER test set as well as two other widely used face detection benchmarks, FDDB and the recently released IJB-A have been presented.

9. *Wei Jang, Wei Wang* - **Face Detection and Recognition for Home Service Robots wth End-To-End Deep Neural Networks, IEEE 2017** [9]
   This paper proposes an effective end-to-end face detection and recognition framework based on deep convolutional neural networks for home

service robots. State-of-the-art region proposal based deep detection network has been combined with he deep face embedding network into an end-to-end system, so that the detection and recognition networks can share the same deep convolutional layers, enabling significant reduction of computation through sharing convolutional features. The detection network is robust to large occlusion, and scale, pose, and lighting variations. The recognition network does not require explicit face alignment, which enables an effective training strategy to generate a unified network. A practical robot system is also developed based on the proposed framework, where the system automatically asks for a minimum level of human supervision when needed, and no complicated region-level face annotation is required. Experiments are conducted over WIDER and LFW benchmarks, as well as a personalized dataset collected from an office setting, which demonstrate state-of-the-art performance of the system.

10. *Rajesh K M, Naveenkumar M* - **A Robust Method for face Recognition and Face Emotion Detection System using Support Vector Machines, IEEE 2016** [10]
    This paper presents framework for real time face recognition and face emotion detection system based on facial features and their actions. The key elements of Face are considered for prediction of face emotions and the user. The variations in each facial feature are used to determine the different emotions of face. Machine learning algorithms are used for recognition and classification of different classes of face emotions by training of different set of images. In this context, by implementing herein algorithms would contribute in several areas of identification, psychological researches and many real world problems. The proposed algorithm is implemented using open source computer vision (OpenCV) and Machine learning with python.

11. *Jie Shen, Ognjen Rudovic, Shiyang Cheng, Maja Pantic* - **Sentiment Apprehension in Human-Robot Interaction with NAO** [11]
    In this paper, the influence of sentiment apprehension by robots (i.e., robot's ability to reason about the user's attitudes such as judgment / liking) on the user engagement has been studied. Two versions of mimicry game are studied: in the first, NAO was solely mimicking facial expressions of the users, while in the second he was also providing a feedback based on the sentiment apprehension. A total of 32 participants (7 female, 25 male) were recruited for this experiment, and

the results show that the participants in the second group spent more time interacting with the robot and played more rounds of the mimicry game. After experiencing both versions of the game, ratings given by the participants indicate (with 99% confidence) that the game with sentiment apprehension is more engaging than the baseline version.

12. *Dario Bertero, Pascale Fung* - **A First Look into Convolutional Neural Network for Speech Emotion Detection, IEEE 2017** [12]
A real-time Convolutional Neural Network model for speech emotion detection. Our model is trained from raw audio on a small dataset of TED talks speech data, manually annotated into three emotion classes: "Angry", "Happy" and "Sad". It achieves an average accuracy of 66.1%, 5% higher than a feature-based SVM baseline, with an evaluation time of few hundred milliseconds. An in-depth model visualization and analysis is also provided. How the neural network effectively activates during the speech sections of the waveform regardless of the emotion, ignoring the silence parts which do not contain information has also been shown. On the frequency domain the CNN filters distribute throughout all the spectrum range, with higher concentration around the average pitch range related to that emotion. Each filter also activates at multiple frequency intervals, presumably due to the additional contribution of amplitude-related feature learning.n

13. *Kun Han, Dong Yu, Ivan Tashev* - **Speech Emotion Recognition and Extreme Learning Machine, INTERSPEECH 2014** [13]
Speech emotion recognition is a challenging problem partly because it is unclear what features are effective for the task. In this paper an approach is proposed to utilize deep neural networks (DNNs) to extract high level features from raw data and it is shown that they are effective for speech emotion recognition. First an emotion state probability distribution is produced for each speech segment using DNNs. Then utterance-level features from segment-level probability distributions are constructed. These utterancelevel features are then fed into an extreme learning machine (ELM), a special simple and efficient single-hidden-layer neural network, to identify utterance-level emotions. The experimental results demonstrate that the proposed approach effectively learns emotional information from low-level features and leads to 20% relative accuracy improvement compared to the stateof-the-art approaches.

14. *Dan Duncan, Gautam Shine, Chris English* - **Facial Emotion Recognition in Real Time** [14]
    This paper proposes a convolutional neural network for classifying human emotions from dynamic facial expressions in real time. Transfer learning is used on the fully connected layers of an existing convolutional neural network which was pretrained for human emotion classification. A variety of datasets and homebrewed dataset is used to train the model. Overall training accuracy of 90.7% and test accuracy of 57.1% is achieved. A live video stream connected to a face detector feeds images to neural network. The network subsequently classifies an arbitrary number of faces per image simultaneously in real time. This paper essentially demonstrates the feasibility of implementing neural networks in real time to detect human emotions.

15. *Lifeng Shang, Zhengdong Lu, Hang Li* - **Neural Responding Machine for Short-Text Conversation** [15]
    This paper proposes Neural Responding Machine (NRM), a neural network-based response generator for Short-Text Conversation. NRM takes the general encoderdecoder framework: it formalizes the generation of response as a decoding process based on the latent representation of the input text, while both encoding and decoding are realized with recurrent neural networks (RNN). The NRM is trained with a large amount of one-round conversation data collected from a microblogging service. Empirical study shows that NRM can generate grammatically correct and content-wise appropriate responses to over 75% of the input text, outperforming stateof- the-arts in the same setting, including retrieval-based and SMT-based models(Statistical Machine Translation or a generative model).

16. *Joost Broekens* - **Emotion and Reinforcement: Affective Facial Expressions Facilitate Robot Learning** [16]
    Computer models can be used to investigate the role of emotion in learning. Here weThis paper presents EARL framework for the systematic study of the relation between emotion, adaptation and reinforcement learning (RL). EARL enables the study of communicated affect as reinforcement to the robot. In humans, emotions are crucial to learning. For example, a parent—observing a child—uses emotional expression to encourage or discourage specific behaviors. Emotional expression can therefore be a reinforcement signal to a child. We hypothesize that affective facial expressions facilitate robot learning, and

compare a social setting with a non-social one to test this. The non-social setting consists of a simulated robot that learns to solve a typical RL task in a continuous grid-world environment. The social setting additionally consists of a human (parent) observing the simulated robot (child). The human's emotional expressions are analyzed in real time and converted to an additional reinforcement signal used by the robot; positive expressions result in reward, negative expressions in punishment. It is quantitatively shown that the "social robot" indeed learns to solve its task significantly faster than its "non-social sibling". This paper concludes that this presents strong evidence for the potential benefit of affective communication with humans in the reinforcement learning loop.

Literature Gap

Emotional states are not clearly defined with boundaries In general literature available today, numerous features have been developed, however the performance of classifiers is still limited, which is because of the fact that emotional states cannot be accurately distinguished by a well-defined set of discriminating features.

Research on tone analysis Also, majority of work done towards emotion detection is focused on a single mode i.e. audio/ text/ video. There is limited practical work done with multimodal inputs.

Response generation systems Response Generation Systems are mostly retrieval based, rather than a completely generative model.

3 Problem Definition

- Knowing the emotional state of an individual can be crucial in determining what action is to be taken as a response.
- Recognizing the affective state of a human can be difficult for humans as well as computer systems. Many features can be considered such as voice samples, facial cues or even text written by the person to identify the emotional state of the individual.
- The major focus of the project is improving human-machine interaction using the NAO robot.
- The robot will accept the input from the person periodically in the form of speech samples, comprising of voice and text as well as facial cues and will interpret the current emotional state of the person.
- Although our main focus is on humanizing the NAO robot and making it an ideal companion for old people, there are myriad of other uses that can be achieved; some of which are:
  - Development of an affect-aware city
  - Add security layer at public venues to detect malicious intent and deal with hostage situation effectively
  - Measure response and ratings in focus groups (consumer response to commercials etc)
  - Wearables that help autistics discern emotion etc.

4 Scope of the problem

- NAO robot will automatically and periodically analyze voice samples and facial cues in order to detect the emotional state of the person interacting with the robot.
- Specified number of frames per second will be analysed for facial cues.
- Audio segments will be analysed via tone for emotion detection.
- Speech text extracted from the audio segments will be aggregated and analysed for emotion.
- The robot will not be able to detect every single complex emotion, but will be limited to a subset of generalized emotions.
- Depending on the emotions and the context of the conversation, the NAO robot will give an appropriate response.
- The response will be a combination of vocal response as well as physical gesture.
- Vocal response generation will be retrieval based. The physical gesture will be calculated from an inbuilt library.
- This humane response will make the robot an ideal companion for old people.

5 Representation of Emotions

1. Dimensional Emotion Space Emotional state is represented with real values in the emotion space along three axes: Valence, Activation and Dominance.

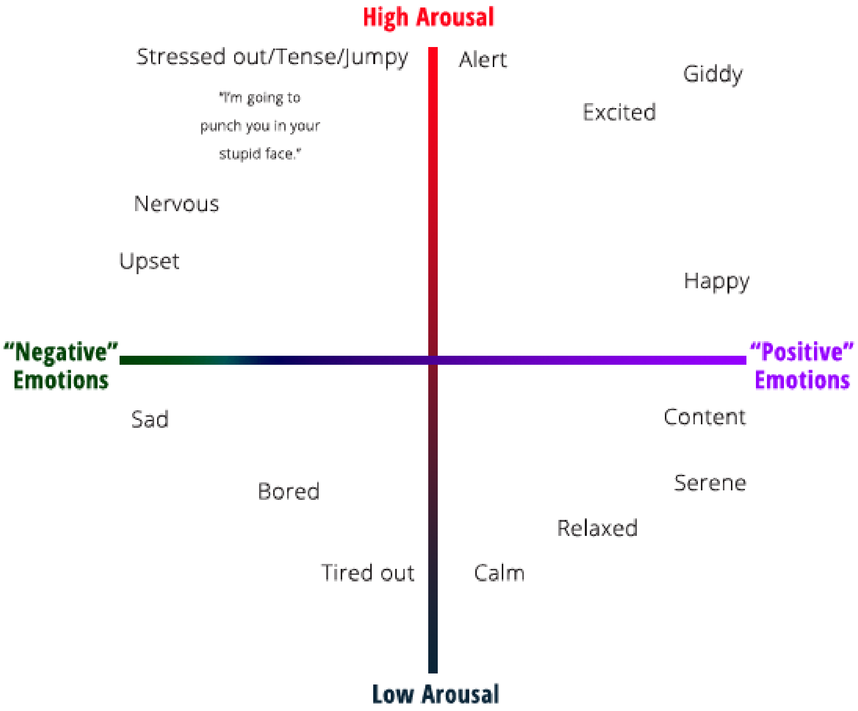

Figure 1: 2-d Emotion Space

   - Valence - Positive vs Negative emotions
   - Activation or Arousal - Active or Passive Emotions.

2. Categorical - The labels refer to one of the basic emotions. The Ekman and Plutchik scales are based on categorical emotion labels.

   - **Ekman scale**
     Paul Ekman determined that there are six basic emotions that are expressed by certain facial expressions that are shared by people in

all cultures. The six basic emotions are anger, happiness, surprise, disgust, sadness, and fear.

- **Plutchik wheel of emotions**

  Robert Plutchik suggested 8 primary bipolar emotions: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation. Additionally, his circumplex model makes connections between the idea of an emotion circle and a color wheel. Like colors, primary emotions can be expressed at different intensities and can mix with one another to form different emotions.

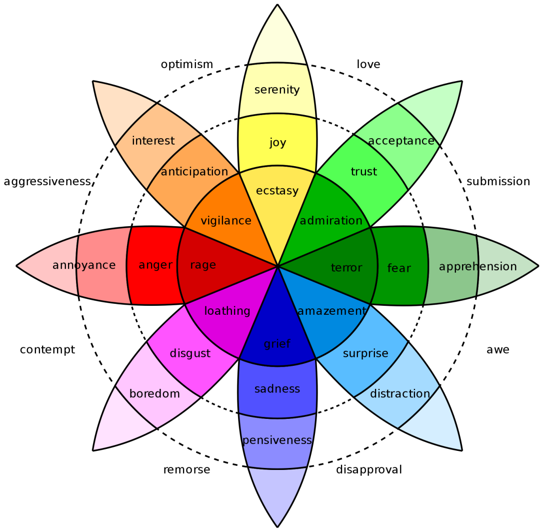

Figure 2: Plutchik wheel of emotions

Robert Plutchik also created a wheel of emotions. This wheel is used to illustrate different emotions in a compelling and nuanced way. Plutchik first proposed his cone-shaped model (3D) or the wheel model (2D) in 1980 to describe how emotions were related.

6 Study of existing projects :

6.1 Emotion Recognition with facial detection

Emotive analytics : blend of psychology and technology. Although reductive, emotions clubbed into 7 main categories : Joy, Sadness, Anger, Fear, Surprise, Contempt, Disgust. For facial emotion detection, algorithms detect faces within photo or video, sense micro-expressions by analyzing the relationship between points on the face, based on curated databases compiled in aacdemic environments. **Sentiment analysis** processing software can analyze text to conclude if a given statement is generally positive or negative based on keywords and their valence index. **Sonic algorithms** analyze recorded speech for both tone and word content

Affectiva Solution for massive scale engagement. SDKs and APIs offered for mobile developers.

Project Oxford by Microsoft Catalogue of artificial intelligence APIs focussed on computer vision, speech and language analysis. Demo takes a photo as an input and output is given in the form of JSON file, with detected faces and emotions of each, as a score between 0 to 1 for each of 8 emotions : anger, contempt, disgust, fear, happiness, neutral, sadness and surprise.

Imotions Biometric research platform providing software and hardware for monitoring many types of bodily cues. Imotion syncs with **Emotient's facial expression technology and adds extra layers to detect confusion and frustration.** Imotions API can monitor video live feeds to extrat valent, or can aggregate previously recorded videos to analyze for emotions. **Used by Harvard, Procter and Gamble, Yale, US Air Force**

CrowdEmotion API that uses facial recognition to detect the time series of the six universel emotions defined by Psychologist Paul Ekman (happniess, surprise, anger, disgust, fear and sadness). Analyses facial points in real-time video and respond with detailed visualizations.

FacioMetrics Founded at Carnegie Mellon University(CMU), provides SDKs for incorporating face tracking, pose and gaze tracking, and expression analysis. Can be tested using **Intraface iOS app**.

6.2 Text to Emotion

Sentiment analysis APIs that provide categorization or entity extraction. Following APIs specifically respond with an emotional summary given a body of plain text.

Natural Language Processing use of machines to detect "natural" human interaction

Deep Linguistic Analysis examination of sentence structure, and relationship between keywords to derive sentiment

IBM Watson Powered by supercomputer IBM Watson, Tone Analyzer detects emotions tones, social propensities and writing styles from any length of plain text. **API can be forked on GitHub.** IBM also provides other cognitive computing tools.

AlchemyAPI Determines relevance of keywords and their associated negative/positive connotations to get a sense of attitude or opinion. URL input can be given to recieve a grade of positive, mixed or negative overall sentiment. Overall sentiment evaluation for the document.

Bitext Text Analysis API is deep linguistic analysis tool. Can be used to analyse words, relations, sentences, structures and dependencies to extract bias with sentiment scoring functionality.

Synesketch(opensource) Analyzes text for sentiment, converts emotional tone into visualizations. **Third-party apps constructed with Synesketch to recognize and visualize emotion from Tweets, speech, poetry and more.**

Tone API Quantifies emotional response for given content. Tool takes a body of text and analyzes for emotional breadth, intensity and comparison with other texts. **Possible application as a service for automating in-house research to optimize smart content publishing.**

6.3 Speech to Emotion

Speech recognition APIs are processed by other sentiment analysis APIs listed above, taking into consideration the **inflection of the speech**. Easy-to-consume web API that instantly recognize emotion from recorded voices are relatively rare. (Use cases : monitoring customer support centers, providing dispatch squads automated emotional intelligence)

Good Vibrations Good Vibrations API senses mood from recorded voice. API and SDK use universal biological signals to perform real time analysis of the user's emotion to sense stress, pleasure, or **disorder**. **EMOSpeech** is enterprise software to analyze emotion. "Audeering" software detects emotion, tone and gender in recorded voice.

Vokaturi Open Vokaturi SDK computes percent likelihoods for 5 emotive states : neutrality, happiness, sadness, anger and fear. (API has code samples for C and python)

7 System architecture

7.1 High Level Design

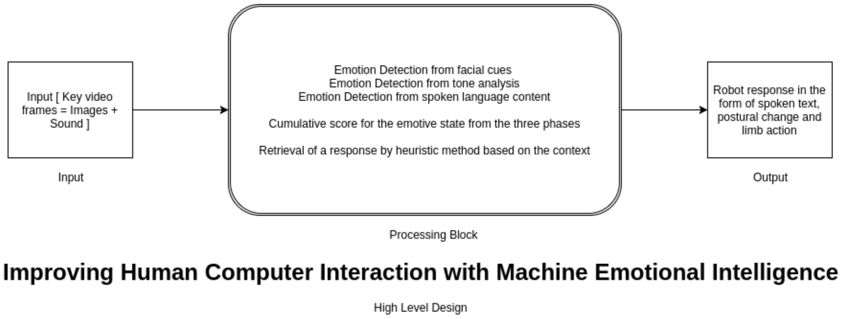
Input [Key video frames = Images + Sound]
Input
Emotion Detection from facial cues
Emotion Detection from tone analysis
Emotion Detection from spoken language content
Cumulative score for the emotive state from the three phases
Retrieval of a response by heuristic method based on the context
Processing Block
Robot response in the form of spoken text, postural change and limb action
Output
Improving Human Computer Interaction with Machine Emotional Intelligence
High Level Design

Figure 3: High Level Design

Input Key video frames and audio segements.

Process Emotion Detection from facial cues, tone analysis, speech text, cumulative emotive state from the three phases, retrieval of a response using a heuristic method.

Output Robot response in the form of speech, postural change and limb action.

7.2 Low Level Design

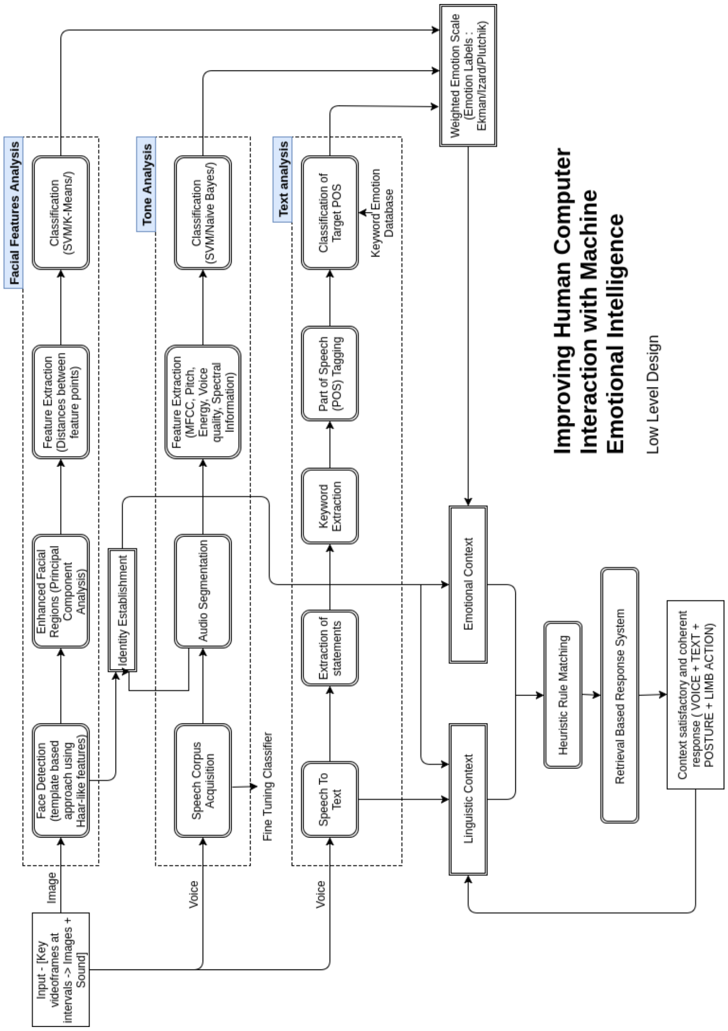
Improving Human Computer Interaction with Machine Emotional Intelligence
Low Level Design
Input - [Key videoframes at intervals -> Images + Sound]
Image
Voice
Voice
Facial Features Analysis
Face Detection (template based approach using Haar-like features)
Enhanced Facial Regions (Principal Component Analysis)
Feature Extraction (Distances between feature points)
Classification (SVM/K-Means/)
Identity Establishment
Tone Analysis
Speech Corpus Acquisition
Audio Segmentation
Feature Extraction (MFCC, Pitch, Energy, Voice quality, Spectral Information)
Classification (SVM/Naive Bayes/)
Fine Tuning Classifier
Text analysis
Speech To Text
Extraction of statements
Keyword Extraction
Part of Speech (POS) Tagging
Classification of Target POS
Keyword Emotion Database
Weighted Emotion Scale (Emotion Labels : Ekman/Izard/Plutchik)
Emotional Context
Linguistic Context
Heuristic Rule Matching
Retrieval Based Response System
Context satisfactory and coherent response (VOICE + TEXT + POSTURE + LIMB ACTION)

Figure 4: Low Level Design

8 Analysis of modules

8.1 Tone analysis

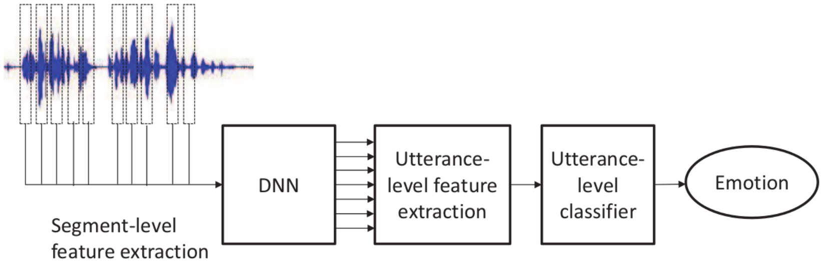
DNN
Utterance-level feature extraction
Utterance-level classifier
Emotion
Segment-level feature extraction

Figure 5: Tone analysis : Mathematical Model - Deep Neural Network(DNN) + Extreme Learning Machine(ELM)

- Input to DNN
  - Frames composed into segments according a sliding window, certain top percentage of segments qualify as input with respect to their energy.
  - MFCC featues, pitch based features are extracted per frame to give feature vector per frame z(m)
  - 2m+1 frames are stacked to generate a segment level feature vector x(m)

$$x(m) = [z(m-w), .., z(m), .., z(m+w)] \tag{1}$$

- Output of DNN
  - A sequence of probability distribution t over all emotion states for each segment

$$t = [P(E_1), ..., P(E_K)]^T \tag{2}$$

- Input to ELM (Utterance level feature extraction)
  Statistical features per each probability distribution.
  f_1, f_2, f_3 which correspond to maximal, minimal and mean of segment-level probabilityof k_{th} level emotion over utterance. f_4 is percentage of the segments which have high probability of emotion *k*.

- Output of ELM(Utterance level classifier)
  K-dimensional vector corresponding to scores of each emotion state. (k emotions considered).

- Objective function for DNN
  Gradient descent - mini-batch

- Objective function for DNN
  Cross entropy

8.2 Speech analysis

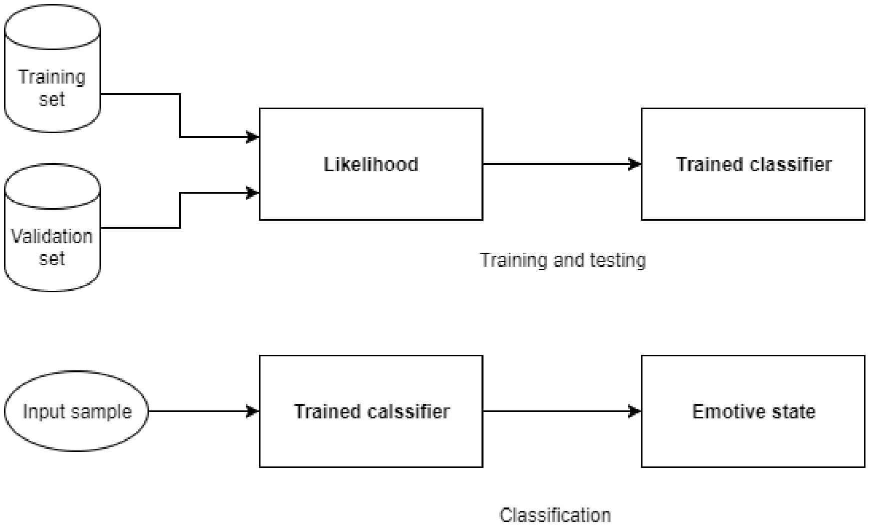

Figure 6: Speech analysis : Mathematical Model - Multinomial Naive Bayesian Classifier

- Training stage
  - Input
    Dataset split into training set and testing(validation) set with k-fold cross-validation for assessing accuracy.
  - Output
    Trained classifier with Likelihood

- – Likelihood(evidence) calculation
  Conditional probabilities of attributes for class labels are calculated from the dataset, termed as evidence Z.

$$Z = p(x) = \sum_k p(C_k)p(x|C_k) \tag{3}$$

- Classification stage
  This stage uses the likelihood or evidence calculated in training to classify a novel input.

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \tag{4}$$

In text classification, the goal is to find the best class for the input text. The best class in Multinomial NB is the most likely or maximum posteriori (MAP) class c_{map}

$$c_{map} = argmax_{c \in C} P(c|d) = argmax_{c \in C} P(c) \prod_{1 \leq k \leq n_d} P(tk|c) \tag{5}$$

In the above equation, many conditional probabilites are multiplied, and with a large enough vocabulary, raw multiplication will almost definitely result in an underflow. It is therefore better to perform the computation by adding logarithms of probabilites instead of multiplying probabilites. The class with highest log probability score is still the most probable; log(xy) = log(x) + log(y) and the logarithmic function is monotonic. Hence, the maximization that is actually done in our implementation of the Multinomial NB classifier is as follows:

$$c_{map} = argmax_{c \in C}[logP(c) + \sum_{1 \leq k \leq n_d} logP(t_k|c)] \tag{6}$$

8.3 Facial feature analysis

- Input
  Image (frame) from the video
- Output
  Prediction based on the output softmax layer
- Process

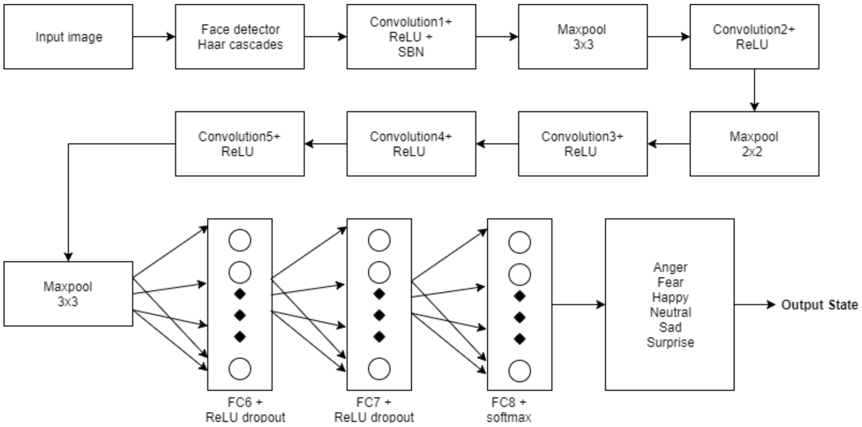

Figure 7: Facial feature analysis : Mathematical Model - Convolutional Neural Network(CNN)

- Image processing is narrowed to the regions of the image containing faces, by performing facial recognition in the image using a classifier (Haars cascades).
- Convolutional Layer
  It computes the output of neurons that are connected to local regions in the input, each computing a dot product between their weights and a small region they are connected to in the input volume.
- ReLU layer
  It applies element-wise activation function which is rectifier function = *max(0,x)* thresholding at zero.
- Pool layer
  It performs downsampling operation along spatial dimensions.
- Fully-Connected (FC) layer
  It computes the class scores among the final categories. Each neuron in this layer is connected to all the numbers in the previous one.
- Dropout randomly ignores nodes to prevent interdependencies emerging between nodes.

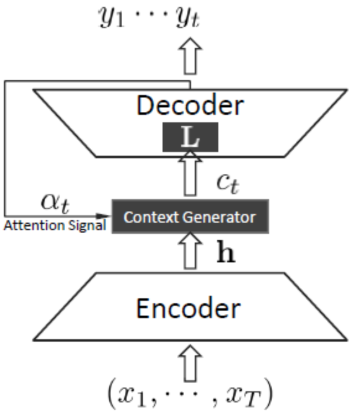

Figure 8: Response generation : Mathematical Model - Neural Responding Machine(NRM)

8.4 Response generation

- Basic idea of NRM is to build a hidden representation of a statement, then generate a response based on it.

- Encoder converts input sequence $(x_1,\ldots,x_T)$ into a set of high-dimensional hidden representations h = $(h_1,\ldots,h_T)$

- h and attention signal (previous decoded response) α_{t} are fed into context generator to build context input to decoder c_t.

- c_t is linearly transformed by matrix L (as a part of the decoder) into a stimulus of generating RNN to produce the t-th word of a response y_t.

9 Hardware and software requirements

9.1 Hardware Requirements

- NAO Robot : Softbank Robotics
  - Height : 58 centimeters
  - Wieght : 4.3 kg
  - Power Supply : Lithium battery providing 48.6 Wh
  - Degrees of freedom : 25
  - Autonomy : 90 minutes (active use)
  - CPU : Intel Atom @ 1.6 Ghz
  - Built-in OS : NAOqi 2.0 (linux-based)
  - Programming Languages : C++, Python, Java, MATLAB, Urbi, C, .NET
  - Sensors : Two HD Cameras, four microphones, sonar rangefinder, two infrared emitters and receivers, intertial board, nine tactile sensors, eight pressure sensors.
  - Connectivity : Ethernet, WiFi
- Server Requirements
  - RAM : 8 GB+ 1333/1600 Mhz
  - CPU : Intel Core (i5/i7 Family)
  - GPU : NVIDIA GPU Accelerator (GeForce Series 9/10 Family)
- Configuration for training classifiers :
  - RAM : 16 GB+ (DDR4 preferred)
  - NVIDIA Tesla GPU Accelerator (K40)
  - Intel Xeon Processor (E5/E7 Family)

9.2 Software Requirements

- Python packages (NPToolKit, python networking libs, etc)
- Continuum packages
- Docker for dependency management

10 Datasets

10.1 Comparitive study of datasets

Table 1: Comparitive study of datasets

Dataset	Source	Modal	Labelling	Language	Comment
Berlin	Acted	Audio	Categorical	German	
eNTERFACE	Acted	Audio + Visual	Categorical	English	
SUSAS	Acted + Actual	Audio	Categorical	English	Stressed scenarios
VAM	Actual	Audio + Visual	Categorical + Dimensional	German	Talkshow
FAUAibo	Acted	Audio + Visual	Categorical	English	Interactive
IEMOCAP	Scripted + Improvised	Audio + Visual + Motion-Capture	Categorical + Dimensional	English	

10.2 IEMOCAP - Interactive Emotional Dyadic Motion Capture database

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is an acted, multimodal and multispeaker database, recently collected at SAIL lab at USC. It contains approximately 12 hours of audiovisual data, including video, speech, motion capture of face, text transcriptions. It consists of dyadic sessions where actors perform improvisations or scripted scenarios, specifically selected to elicit emotional expressions. IEMOCAP database is annotated by multiple annotators into categorical labels, such as anger, happiness, sadness, neutrality, as well as dimensional labels such as valence, activation and dominance. The detailed motion capture information, the interactive setting to elicit authentic emotions, and the size of the database make this corpus a valuable addition to the existing databases in the community for the study and modeling of multimodal and expressive human communication.

10.3 JAFFE - The Japanese Female Facial Expression database

The database contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects. The database was planned and assembled by Michael Lyons, Miyuki Kamachi, and Jiro Gyoba. We thank Reiko Kubota for her help as a research assistant. The photos were taken at the Psychology Department in Kyushu University.

11 Feasibility study

11.1 Hardware feasibility

- NAO robot availabile for 2-3 hours atleast thrice a week.
- External servers to carry out the processing of input and provide outpu.
- Audio and video sensors available in the NAO robot itself.
- Any computer with a GPU can be used to train the machine learning models.

11.2 Software feasibility

- "Choreographe" software and key already available in order to interface with the NAO robot.
- Machine learning libraries necessary for the project are opensource.
- Algorithms required for the project have free to use licenses.
- The datasets used to train the CNN, RNN and SVM are freely available for reference and download

12 Design (UML Diagrams)

12.1 Activity Diagram

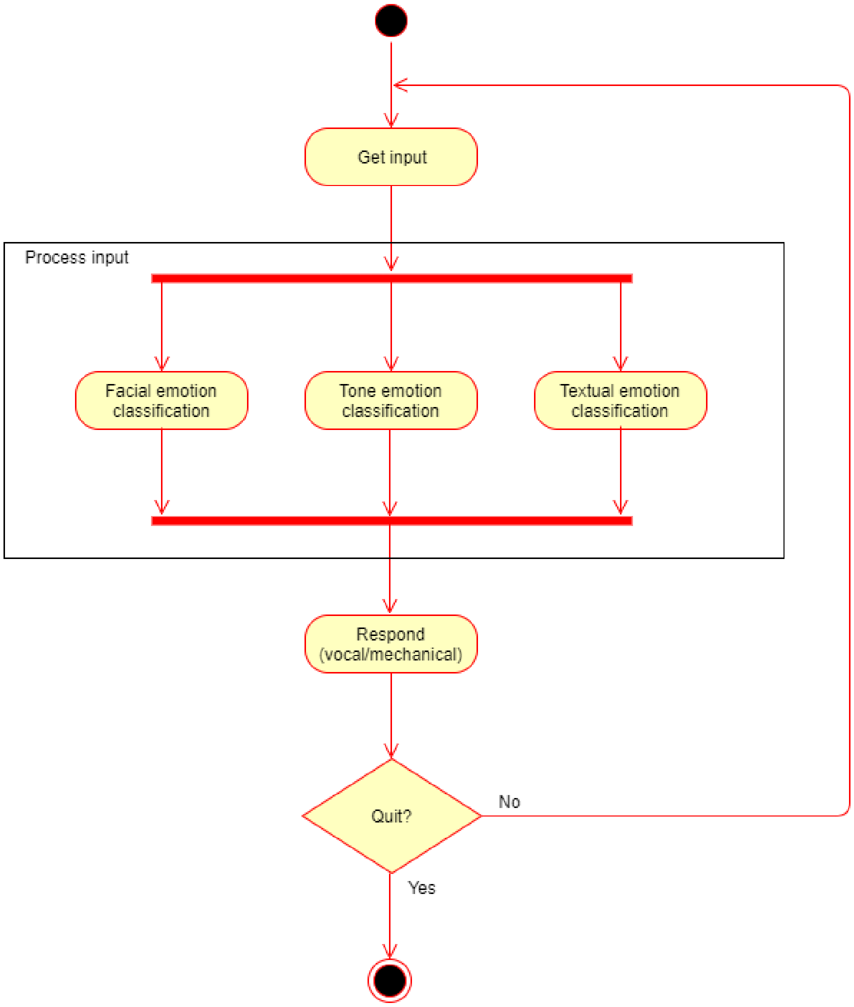
Get input
Process input
Facial emotion classification
Tone emotion classification
Textual emotion classification
Respond (vocal/mechanical)
Quit?
No
Yes

Figure 9: Activity Diagram

12.2 Class Diagram

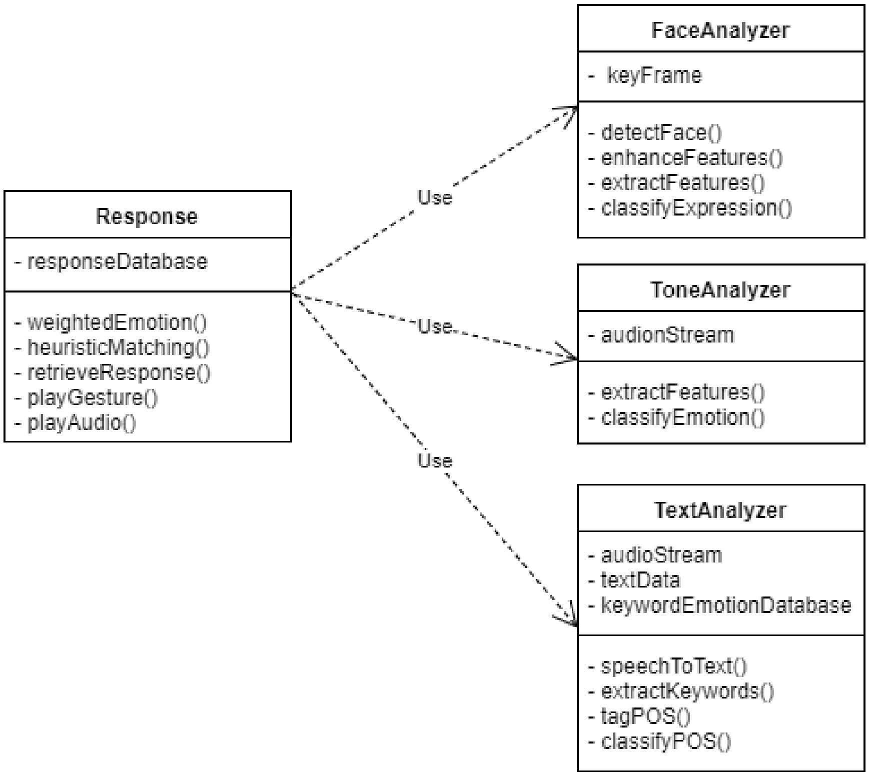

Figure 10: Class Diagram

12.3 Dataflow Diagram

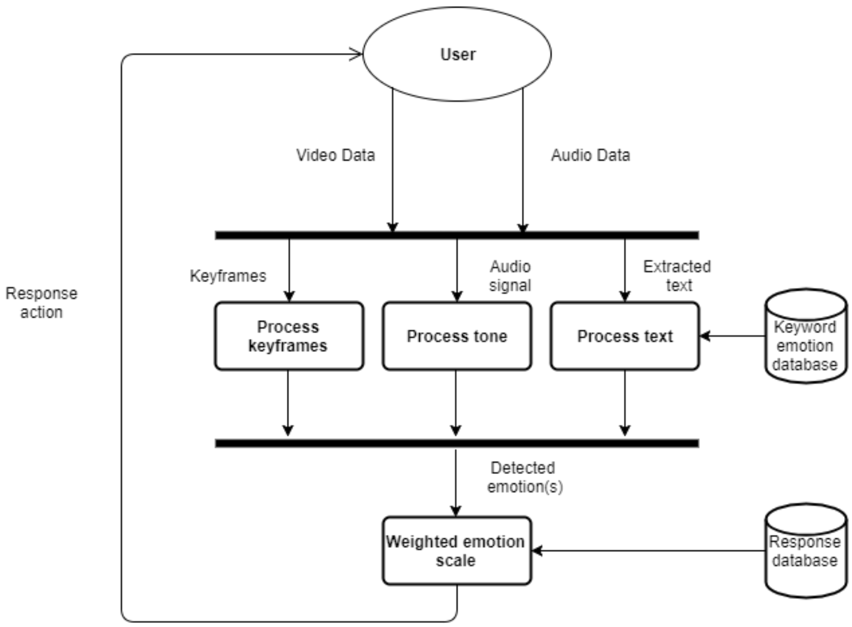

Figure 11: Dataflow Diagram

12.4 Use Case Diagram

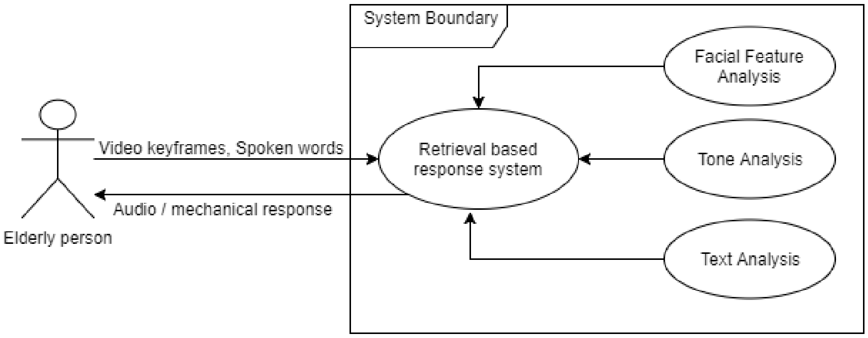
System Boundary
Facial Feature Analysis
Video keyframes, Spoken words
Retrieval based response system
Tone Analysis
Audio / mechanical response
Elderly person
Text Analysis

Figure 12: Use Case Diagram

13 Implementations

13.1 Tone Analysis with DNN and ELM

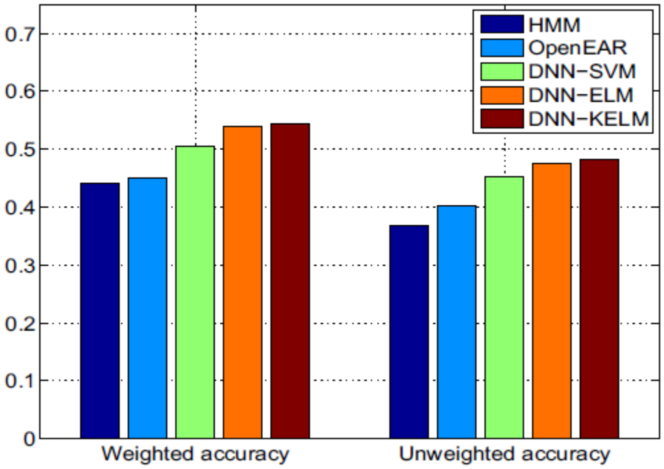

Figure 13: Comparison of DNN and ELM with other approaches

Speech emotion recognition using DNN and ELM [13] has been demonstrated to have the best performance till date, outperforming the previous tonal analysis approaches. Hence, we have implemented a version of the same.

13.1.1 Relevant features for emotion recogntion from voice tone.

- Local features (Low Level Descriptors or LLDs)
  - These features are extracted per frame.
  - Pitch ,magnitude, zero-crossing rate, MFCC are local features.
  - Classifiers used for local features :
    1. GMM : Gaussian mixture models
    2. GMM + UBM : Gaussian mixture models with Universal Background Model
    3. Temporal dynamics : Hidden Markov Model
    4. Latent Dirichlet Association

- Global features (Segment Descriptors)
  - There features are also termed as high level features.
  - They are extracted per utterance.
  - Global features are a combination of a set of local features from each audio frame
  - These are statistical functions for each occurrence.
  - Mean, standard deviation, maximal, minimal, quantiles etc are global features.
  - Classifiers for global features
    1. SVMs
    2. K-Nearest Neighbour
    3. Decision tree
    4. Deep Belief Network
- Decision is based on utterance. Global features or per-utterance features are combined from local features to generate the final ouput.

13.1.2 Approach

1. Local feature extraction
   - The audio is sampled and time domain is converted to frequency domain.
   - A sliding window of 25ms with a 10ms step size processes the given audio sample.
   - A segment stacks 25 frames, 12 past frames, 12 future frames and the current frame. Length of the segment in 265 ms.
   - Features extracted are HNR, MFCC and delta features.
   - Segment feature vector contains frame feature vectors of 25 stacked frames

   $$x(t) = [x(t-12), .., x(t), .., (xt+12)] \tag{7}$$

2. DNN classification
   - Inputs to DNN are segment level feature vectors
   - All segments from same utterance use the same label for training.

- Only the top 10% segments with respect to energy are utilized for training, the rest are discarded which contain silence, noise and speech with no discriminative value for emotion recognition.
- Output from DNN is a probability vector for each emotion for the segment.

3. DNN configuration

    - This DNN is implemented with 1 hidden layer, input layer has 350 units, hidden layer has 500 units.
    - Neurons are sigmoid activated.
    - Objective function used is cross entropy error.
    - Learning algorithm is mini-batch stochastic gradient descent.
    - DNN produces the probabilities for each of the emotion states in the segment.
      $$y(t) = [P(E_1), ..., P(E_k)] \tag{8}$$
      for k=5 emotions = excitement, frustration, happniess, neutral and sadness.

4. Utterance level decision using ELM

    - ELM is a special single hidden layer neural network where input to hidden layer weights are random and hidden layer to output layer weights are trained using least square errors. The number of hidden units is very large compared to the number of input units.
    - Input to ELM is maximum, minimum, mean of $\mathrm{P(E_i)}$ and
      $$|(P(E_i(t)) > \theta| \tag{9}$$
      which is the number of segments for $\mathrm{E_i}$ with probability greater than $\theta = 0.2$, that is, how many segements in the utterance support this emotion.
    - Output of ELM is emotion score vector for entire utterance.
    - Input of ELM has 20 units, 4 high level features (min, man, min, number of supporting segements) per emotion and 5 emotions in total.
    - Number of hidden units is chosen as 120, and weights are assigned randomly.

- To train the connection weights for hidden layer and output layer,

$$min_w||HW - T|| \tag{10}$$

$$W = H^{-1}T \tag{11}$$

  H is the hidden layer to output layer matrix, and T is the training sample.

- Advantages of ELM
  - Since weights are trained by least squared error and not by gradient descent, training is very fast.
  - Since weights are random, the hidden representation is not highly dependent on the training data, thus resulting in good generalisation performance.
  - Observed performance is better than SVM.

5. Performace measures

- Weighted Accuracy

$$WA = |C|/|S| \tag{12}$$

  where C : correctly labelled utterances and S : all utterances

- Unweighted Accuracy The input dataset may be skewed towards particular emotion, mostly neutral, as the annotators are more likely to evaluate an ambiguos utterances as neutral.

  To compensate for this imbalanace in the classes of the training data, we consider unweighted accuracy :

$$UA = 1/N \sum_{i=1}^{N}(|Ci|/|Si|) \tag{13}$$

- Performance measures don't completely reflect the performace of the classifier, as a sample may have more than one connotations, but will be considered a misclassification if it is some other class than the annotated class.

6. Dataset

- IEMOCAP dataset is used.
- It is an acted, multimodal, multispeaker dataset which includes video, speech, motion capture of face and text transcripts of the dialogs.

- Each utterance is annotated by atleast 3 human evaluators including the actor. Annotations are both categorial and dimensional.
- Corpus is created by selecting statements that have been given the same label by atleast 2 annotators.

7. Results Confusion Matrix :

Table 2: Confusion matrix for DNN and ELM : Tone analysis

	exc	fru	hap	neu	sad
exc	0.50	0.14	0.03	0.02	0.05
fru	0.25	0.63	0.19	0.37	0.07
hap	0.01	0.02	0.16	0.05	0.04
neu	0.17	0.14	0.24	0.35	0.10
sad	0.07	0.07	0.38	0.21	0.73
Accuracy	0.50	0.63	0.16	0.35	0.73

13.2 Speech Text Analysis with Multinomial Naive Bayesian Classification

1. Why choose Multinomial Naive Bayes?

- Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of features in a learning problem. As a result, new emotions can be added seamlessly in our emotion recognition model as progress is made.
- Time complexity
  - Training

$$\Theta(|D|L_{avg} + |C||V|) \tag{14}$$

  - Testing

$$\Theta(L_a + |C|M_a) = \Theta(|C|M_a) \tag{15}$$

\| D \| : Total number of documents
L_{avg} : Average length of a document
\| C \| : Total number of classes
\| V \| : Size of the vocabulary
L_a : Number of words in the testing document
M_a : Number of unique words in the testing document

    - Both trainin and testing complexity are linear in time it takes to scan the data. Because we habe to look at the data atleast once, Multinomial NB can be said to have optimal time complexity. Its efficiency is one of the main reasons why NB is chosen as the text classifier for the project.

- Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used by many other classifiers. Some of the alternatives are:

    (a) SVM This a popular alternative, but is not nearly as scalable as our problem demands, since it is limited to binary classification. Furthermore, selection of parameters and kernel functions can be problematic.

    (b) Neural Networks Training is much more slower as compared to Multinomial NB. Word vectorization is needed, in which the dimensionality of vectors increases exponentially with increase in vocabulary.

    (c) KNN Computational cost is comparatively very high. The classification time is also much longer. Finding the optimal value of ‘k’ is another problem, and it may differ for various inputs of different sizes.

2. Pre processing We used the python library NLTK for performing preprocessing tasks. Initially, an input text is given as an input to a word tokenizer which converts the text into individual word tokens. After this, POS tagging is performed on the converted tokens, which labels the tokens as different parts of speech. Some of the labels used by the NLTK POS tagger are as follows:
   CC - coordinating conjunction
   CD - cardinal digit
   DT - determiner
   JJ - adjective
   NN - noun
   RB - adverb
   VB - verb
   UH - interjection
   IN - preposition
   We assessed that most of these labels are irrelevant to the task of emotion recognition. For instance, use prepositions such as ‘on’, ‘at’,

‘in’, etc in text is highly unlikely to be affected by the person’s emotion. From our analysis, we found nouns, verbs, adverbs, and adjectives to be the most useful labels for the emotion recognition task. To filter out the remaining labels, we used the following regular expression to match the labels of the tokens:
NN[A-Z]* | RB[A-Z]* | JJ[A-Z]* | VB[A-Z]*

3. Training the classifier Very little work has been done in emotion recognition from spoken text. Most of the work in the field has been done on written text, such as tweets, and data sets exist for the same. However, there is almost no data for training based on spoken text, which is what our model is based on. So, we are attempting to use the features from written data such as reviews and Twitter tweets, that are likely to occur in spoken text. We do this by removing useless symbols such as ‘#’, ‘@’, and other words that occur only in written text, such as abbreviations an interjections.

4. Datasets

   (a) EmoBank 10,000 sentences annotated with Valence, Arousal and Dominance values.

   (b) Emotion in Text dataset by CrowdFlower 40,000 tweets from Twitter manually classified into 14 emotions such as anger, worry, happiness, fun, sadness, etc.

   (c) Dataset by Sanders 5513 hand-classified tweets, each classified with respect to one of four topics – Positive, Neutral, Negative and Irrelevant.

14 DONE Time-line analysis of the project

Table 3: Time line analysis

Time	Tasks

15 Future scope

Multimodal inputs from multiple channels Algorithms currently designed assume that the signal per module is single channel. They can be extended to include multiple channels for a better affect-awareness module.

Using normalization per subject The models described assume that no information about the subject is available prior to exposure of the model to the subject. The collected information over a period of time can be used for user adaptation and personalization, for example, to implement gesture dynamics.

Facial gaze tracking Along with facial features, gaze tracking can be implemented to improve emotion recognition.

Heirarchical training In the tone analysis module, 10% of the samples with highest energy are utilized for training DNN. Heirarchical learning can be implemented to choose best training samples directly. (Apply training sub-datasets to a sequence of DNNs, during testing stage, the input sample is simultaneously applied to all of these DNNs and decision is aggregated).

Incorporate temporal dynamics Speed at which a transition can take place is termed as temporal dynamics of a system. The change in emotional state can be studied and incorporated into the recognition mechanism.

Learning from spectrograms Instead of using handcrafted feaures like MFCC, it should be possible to learn emotional information from spectrogram with techniques like Mel-filter bank.

16 Conclusion

We proposed an emotion recognition system comprising of three modules, tone, speech text and facial feature analysis. We tested the implementation of tone analysis model and speech text analysis model for emotion recognition with promising experimental results. The output emotive state of this emotion recognition system is to be used as reinforcement to determine a response using a neural responding machine. Thus implementing an emotion based reinforcement loop will make interaction with a robot more humane.

References

[1] Jeong-Sik Park and Gil-Jin Jang. Implementation of voice emotion recognition for interaction with mobile agent. *HAI*, 2015.

[2] Yu Gu, Eric Postma, and Hai-Xiang Lin. Vocal emotion recognition with log-gabor filters. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge - AVEC '15*, page nil, - 2015.

[3] Gloria Zen, Enver Sangineto, Elisa Ricci, and Nicu Sebe. Unsupervised domain adaptation for personalized facial emotion recognition. In *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, page nil, - 2014.

[4] Ahmed Mustafa Mahmoud and Wan Haslina Hassan. Determinism in speech pitch relation to emotion. In *Proceedings of the 2nd International Conference on Interaction Sciences Information Technology, Culture and Human - ICIS '09*, page nil, - 2009.

[5] Nancy Semwal, Abhijeet Kumar, and Sakthivel Narayanan. Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, page nil, 2 2017.

[6] Lei Pang and Chong-Wah Ngo. Mutlimodal learning with deep boltzmann machine for emotion prediction in user generated videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval - ICMR '15*, page nil, - 2015.

[7] Benjamin Guthier, Rajwa Alharthi, Rana Abaalkhail, and Abdulmotaleb El Saddik. Detection and visualization of emotions in an affect-aware city. In *Proceedings of the 1st International Workshop on Emerging Multimedia Applications and Services for Smart Cities - EMASC '14*, page nil, - 2014.

[8] Huaizu Jiang and Erik Learned-Miller. Face detection with the faster r-cnn. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, page nil, 5 2017.

[9] Wei Jiang and Wei Wang. Face detection and recognition for home service robots with end-to-end deep neural networks. In *2017 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), page nil, 3 2017.

[10] K. M. Rajesh and M. Naveenkumar. A robust method for face recognition and face emotion detection system using support vector machines. In *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*, page nil, 12 2016.

[11] Jie Shen, Ognjen Rudovic, Shiyang Cheng, and Maja Pantic. Sentiment apprehension in human-robot interaction with nao. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, page nil, 9 2015.

[12] Dario Bertero and Pascale Fung. A first look into a convolutional neural network for speech emotion detection. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page nil, 3 2017.

[13] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. *Interspeech*, 2014.

[14] Dan Duncan, Gautam Shine, and Chris English. Facial emotion recognition in real time. *Stanford*, 2016.

[15] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, page nil, - 2015.

[16] Joost Broekens. Emotion and reinforcement: affective facial expressions facilitate robot learning. In *Artificial Intelligence for Human Computing*, pages 113–132, 2007.