Project Synopsis

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

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

Acknowledgement

Contents

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 Adaptation 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. *Yu Gu, Eric Postma, Hai-Xing Lin* - **Vocal Emotion 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.

12. *Lei Pang, Chong-Wah Ngo* - **Multimodal Learning with Depp 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.

13. *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.

14. *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 evalua-

tion 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.

15. *Gloria Zen, Elisa Ricci, Nicu Sebe* - **Unsupervised Domain Adaptation 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.

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 System architecture

5.1 High Level Design

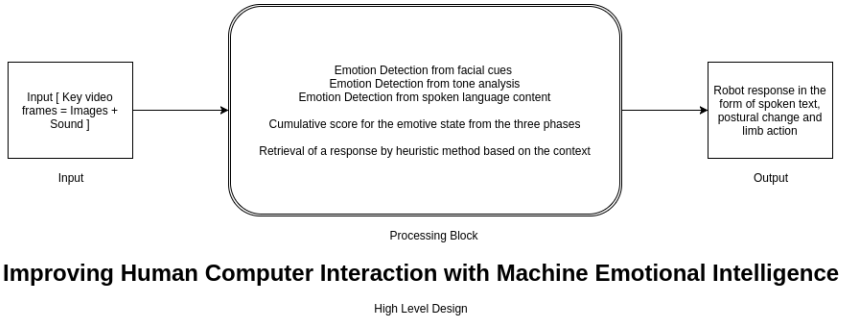

Figure 1: High Level Design

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.

5.2 Low Level Design

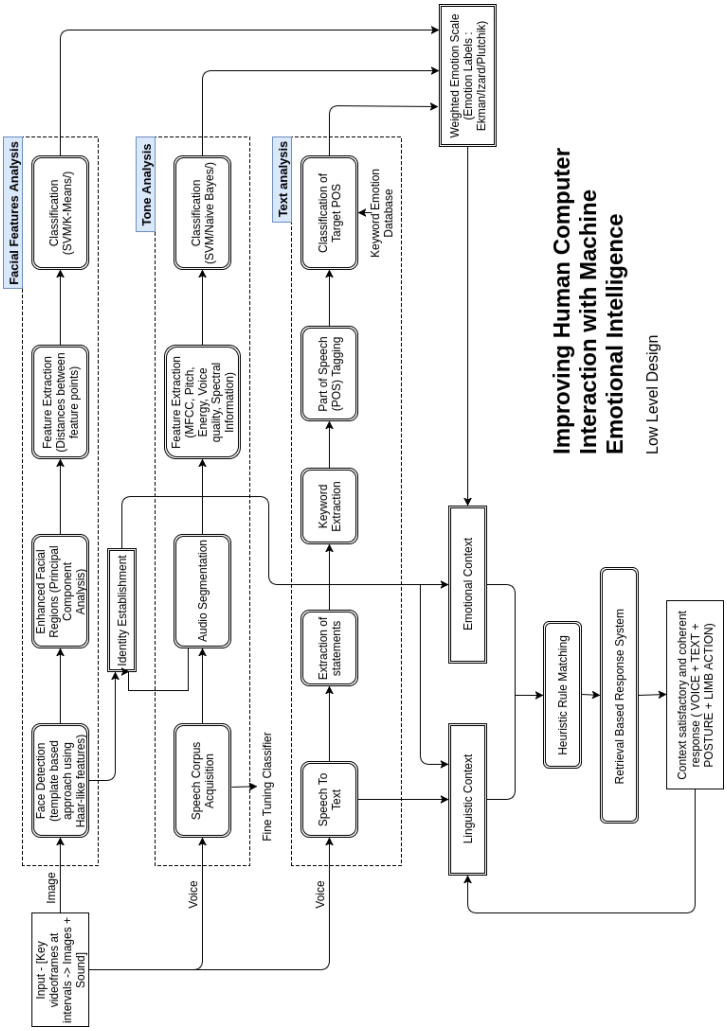

Figure 2: Low Level Design

5.3 Tone analysis : RNN

5.3.1 Recurrent Neural Networks

A recurrent neural network (RNN) is a class of artificial neural network where connections between units form a directed cycle. This allows it to exhibit dynamic temporal behavior. Unlike feedforward neural networks, RNNs can use their internal memory to process arbitrary sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. RNNs come in many variants:

1. Fully recurrent Basic RNNs are a network of neuron-like nodes, each with a directed (one-way) connection to every other node.[citation needed] Each node (neuron) has a time-varying real-valued activation. Each connection (synapse) has a modifiable real-valued weight. Nodes are either input nodes (receiving data from outside the network), output nodes (yielding results), or hidden nodes (that modify the data en route from input to output).
   For supervised learning in discrete time settings, sequences of real-valued input vectors arrive at the input nodes, one vector at a time. At any given time step, each non-input unit computes its current activation (result) as a nonlinear function of the weighted sum of the activations of all units that connect to it. Supervisor-given target activations can be supplied for some output units at certain time steps. For example, if the input sequence is a speech signal corresponding to a spoken digit, the final target output at the end of the sequence may be a label classifying the digit.
   In reinforcement learning settings, no teacher provides target signals. Instead a fitness function or reward function is occasionally used to evaluate the RNN's performance, which influences its input stream through output units connected to actuators that affect the environment. This might be used to play a game in which progress is measured with the number of points won.
   Each sequence produces an error as the sum of the deviations of all target signals from the corresponding activations computed by the network. For a training set of numerous sequences, the total error is the sum of the errors of all individual sequences.

2. Recursive A recursive neural network is created by applying the same set of weights recursively over a differentiable graph-like structure by traversing the structure in topological order. Such networks are typically also trained by the reverse mode of automatic differentiation.

They can process distributed representations of structure, such as logical terms. A special case of recursive neural networks is the RNN whose structure corresponds to a linear chain. Recursive neural networks have been applied to natural language processing. The Recursive Neural Tensor Network uses a tensor-based composition function for all nodes in the tree.

3. Neural history compressor The neural history compressor is an unsupervised stack of RNNs. At the input level, it learns to predict its next input from the previous inputs. Only unpredictable inputs of some RNN in the hierarchy become inputs to the next higher level RNN, which therefore recomputes its internal state only rarely. Each higher level RNN thus studies a compressed representation of the information in the RNN below. This is done such that the input sequence can be precisely reconstructed from the representation at the highest level.
   The system effectively minimises the description length or the negative logarithm of the probability of the data. Given a lot of learnable predictability in the incoming data sequence, the highest level RNN can use supervised learning to easily classify even deep sequences with long intervals between important events.
   It is possible to distill the RNN hierarchy into two RNNs: the "conscious" chunker (higher level) and the "subconscious" automatizer (lower level). Once the chunker has learned to predict and compress inputs that are unpredictable by the automatizer, then the automatizer can be forced in the next learning phase to predict or imitate through additional units the hidden units of the more slowly changing chunker. This makes it easy for the automatizer to learn appropriate, rarely changing memories across long intervals. In turn this helps the automatizer to make many of its once unpredictable inputs predictable, such that the chunker can focus on the remaining unpredictable events.

4. Second order RNNs Second order RNNs use higher order weights instead of the standard weights, and inputs and states can be a product. This allows a direct mapping to a finite state machine both in training, stability, and representation. Long short-term memory is an example of this but has no such formal mappings or proof of stability.

5. Recurrent multilayer perceptron network Generally, a Recurrent Multi-Layer Perceptron (RMLP) network consists of cascaded subnetworks, each of which contains multiple layers of nodes. Each of these subnetworks is feed-forward except for the last layer, which can have feedback

connections. Each of these subnets is connected only by feed forward connections.

5.3.2 Training the RNN

Gradient Descent Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. In neural networks, it can be used to minimize the error term by changing each weight in proportion to the derivative of the error with respect to that weight, provided the non-linear activation functions are differentiable.
The standard method is called "backpropagation through time" or BPTT, and is a generalization of back-propagation for feed-forward networks. Like that method, it is an instance of automatic differentiation in the reverse accumulation mode of Pontryagin's minimum principle. A more computationally expensive online variant is called "Real-Time Recurrent Learning" or RTRL,which is an instance of automatic differentiation in the forward accumulation mode with stacked tangent vectors. Unlike BPTT, this algorithm is local in time but not local in space.
A major problem with gradient descent for standard RNN architectures is that error gradients vanish exponentially quickly with the size of the time lag between important events.

5.4 Facial Feature analysis : CNN

5.4.1 Convolutional Neural Network

In machine learning, a convolutional neural network (CNN, or ConvNet) is a class of deep, feed-forward artificial neural networks that has successfully been applied to analyzing visual imagery. CNNs use a variation of multilayer perceptrons designed to require minimal preprocessing. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics.
Convolutional networks were inspired by biological processesin which the connectivity pattern between neurons is inspired by the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field.
CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in

traditional algorithms were hand-engineered. This independence from prior knowledge and human effort in feature design is a major advantage. They have applications in image and video recognition, recommender systems and natural language processing.

5.4.2 The convolutional layers

The convolutional layer is the core building block of a CNN. The layer's parameters consist of a set of learnable filters (or kernels), which have a small receptive field, but extend through the full depth of the input volume. During the forward pass, each filter is convolved across the width and height of the input volume, computing the dot product between the entries of the filter and the input and producing a 2-dimensional activation map of that filter. As a result, the network learns filters that activate when it detects some specific type of feature at some spatial position in the input.
Stacking the activation maps for all filters along the depth dimension forms the full output volume of the convolution layer. Every entry in the output volume can thus also be interpreted as an output of a neuron that looks at a small region in the input and shares parameters with neurons in the same activation map.

5.5 Speech Text analysis : SVM

- In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

- In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

- When data are not labeled, supervised learning is not possible, and

an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering and is often[citation needed] used in industrial applications either when data are not labeled or when only some data are labeled as a preprocessing for a classification pass.

6 Hardware and software requirements

6.1 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)

6.2 Software Requirements

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

6.3 Datasets

6.3.1 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.

6.3.2 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.

7 Feasibility study

7.1 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.

7.2 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

7.3 Study of existing projects :

7.3.1 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

Emotient For adverisement campaigns that want to track attention, engagement and sentiment from viewers. Provide RESTful Emotient Web API.

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

EmoVu Facial detection product incorporates machine learning and micro-expression detection that allow accurate measurement of content's emotional engagement and effectiveness on their target audience. Desktop SDK, Mobile SDK and API for fine grained control provided by Eyeris. Other features offered by the platform are head position, tilt, eye tracking, eye open/close etc.

Nviso Specialise in emotion video analytics, using 3d facial imaging tech to monitor many different facial data points to produce likelihood for 7 main emotions. Aearder for smarter computing in 2013 by IBM.

Kairos Emotion Analysis API as Saas, coordinates detected in the input video that represent smiles, surprise, anger, dislike and drowsiness.

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.

Face Reader by Noldus Used in academic sphere, Face Reader API is based on machine learning, dataset of 10,000 facial expression images. API uses 500 key facial points to analyze 6 basic facial expressions to analyse emotions, as well as gaze direction and head orientation.

SightCorp Facial Recognition Provider. Insight SDK tracks hundred of facial points, eye gaze tested in museum showcases and at TEDx Amsterdam.

SkyBiometry Cloud based face detection and recognition tool for detecting emotion in photos. Output is a percentage rate for moods : happy, sad, angry, surprised, disgusted, scared and neutral, in a given photo input.

Face++ Facial recognition tool that compares faces with stored faces, targeted for name tagging in photos in social networks. Determines if face is smiling or not. Provides a set of developer SDKs.

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**.

7.3.2 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.

Receptiviti Natural Language Personality Analytics API uses a process of target words and emotive categories to derive an emotion and personality from texts. Their Linguistic Inquiry and Word Count (LIWC) text analysis process used by IBM. Provides endpoints for REST API and SDKs in all major languages.

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.

Mood Patrol Hosted on Mashape API marketplace, extracts emotions from text. It responds with fine grained adjectives that describe emotional tone based on Plutchik's 8 Basic Emotions.

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.**

Repustate API Repustate Sentiment Analysis process is based in linguistic theory, reviews cues from lemmatization, polarity, negations and parts of speech and more to reach informed sentiment from a text document.

7.3.3 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)

8 Design (UML Diagrams)

8.1 Activity Diagram

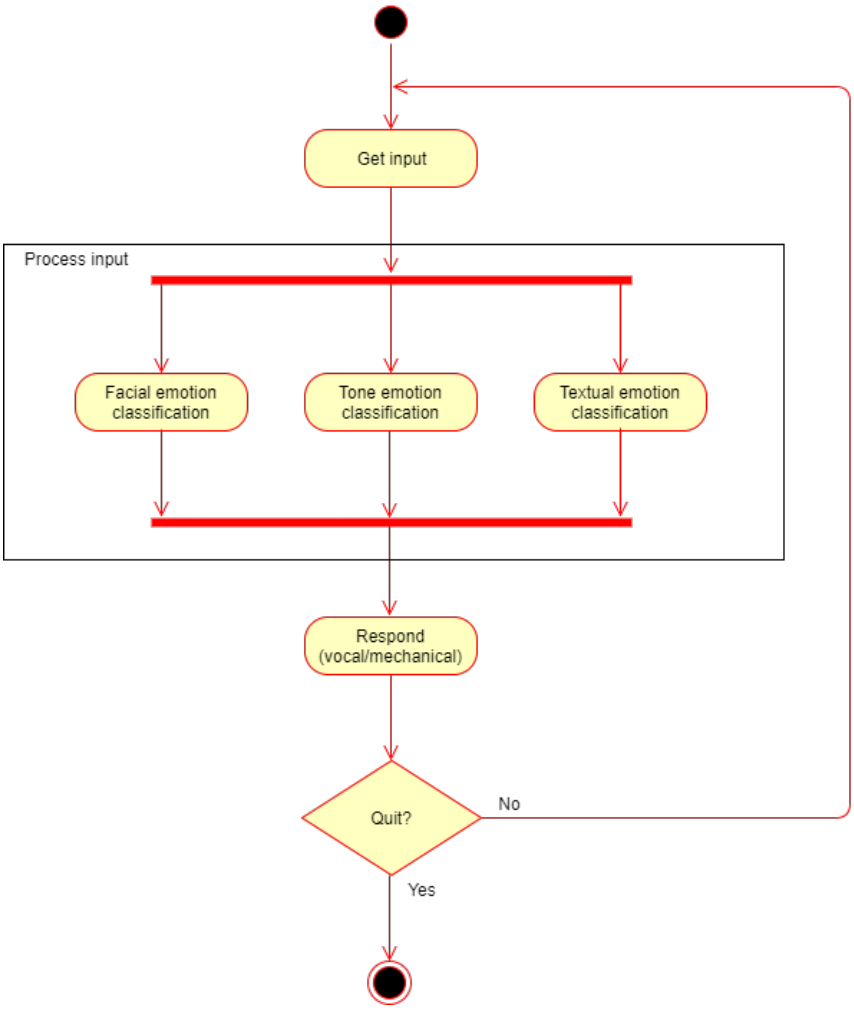

Figure 3: Activity Diagram

8.2 Class Diagram

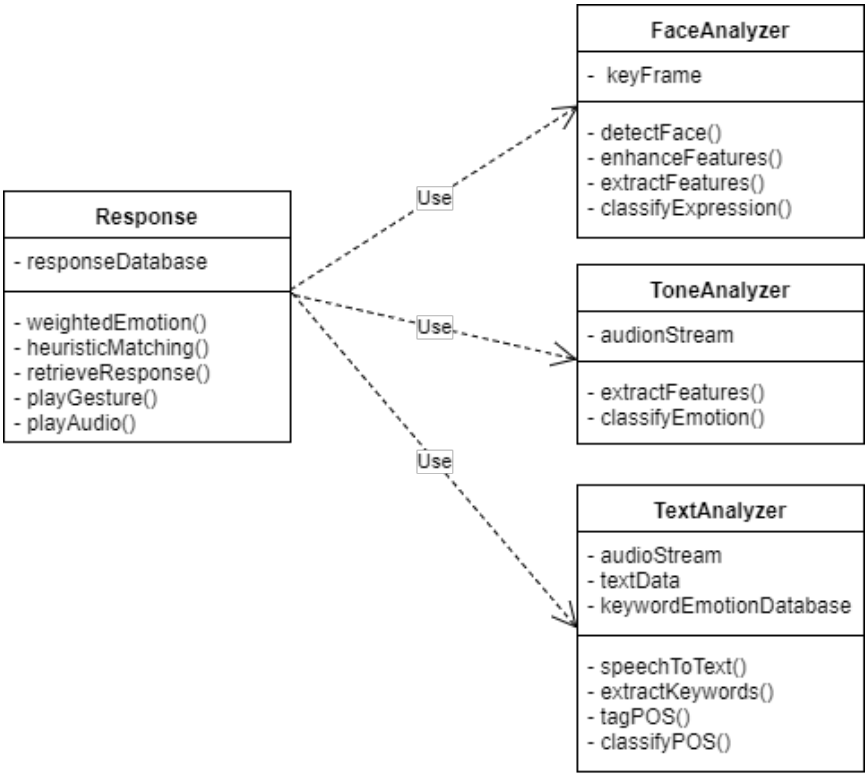

Figure 4: Class Diagram

8.3 Dataflow Diagram

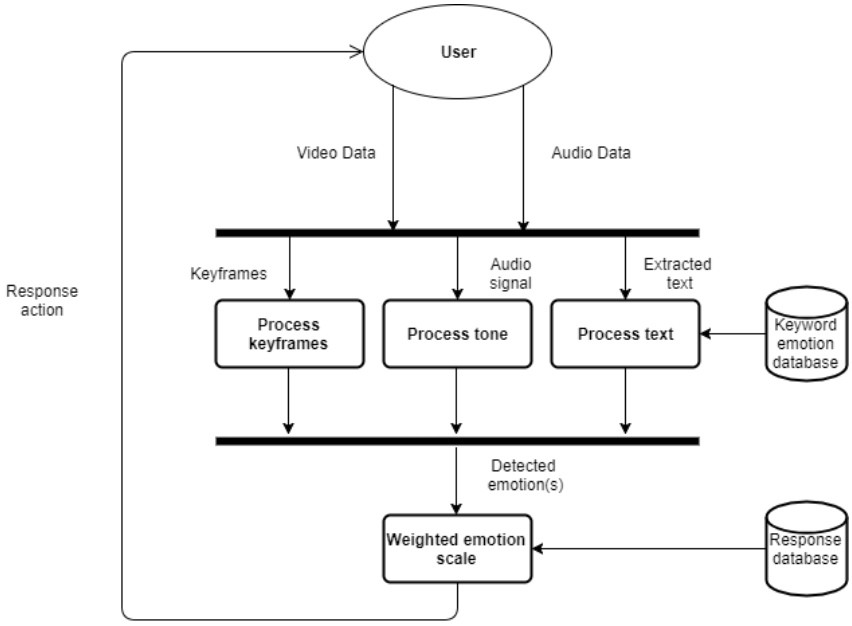

Figure 5: Dataflow Diagram

8.4 State Diagram

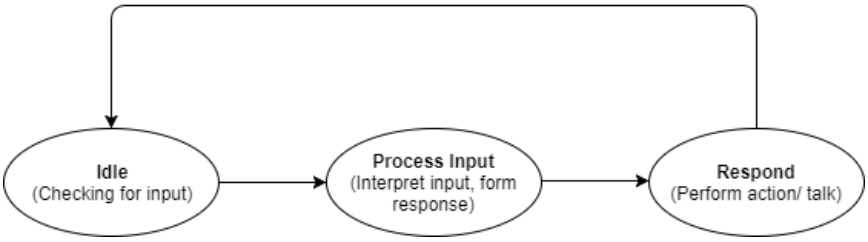

Figure 6: State Diagram

8.5 Use Case Diagram

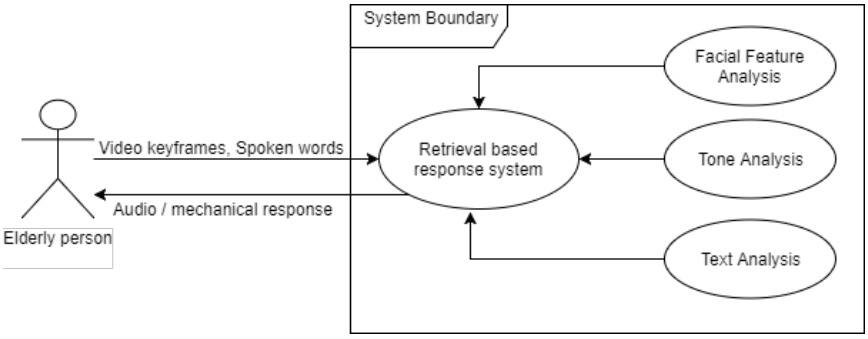

Figure 7: Use Case Diagram

9 TODO Time-line analysis of the project

10 TODO Future scope

11 TODO Conclusion

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.