Experience Based Adaptive Vocal Interaction System for Domestic Service Robots

Sajila Dimagi Wickramaratne

(168077J)

Thesis submitted in partial fulfillment of the requirements for the degree Master of Science in Electrical Engineering

Department of Electrical Engineering

University of Moratuwa

Sri Lanka

June 2018

Experience Based Adaptive Vocal Interaction System for Domestic Service Robots

Sajila Dimagi Wickramaratne

(168077J)

Thesis submitted in partial fulfillment of the requirements for the degree Master of Science in Electrical Engineering

Department of Electrical Engineering

University of Moratuwa

Sri Lanka

June 2018

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the MSc thesis under my supervision.

Signature of the Supervisor(s): Date:

Dr.A.G.B.P.Jayasekara

Abstract

Domestic service robots have gained an increasing popularity in the recent times due to the lack of personnel to tend to domestic work. Domestic service robots with conversational capabilities can be sought out as an exceptional solution to eliminate the mental distress among the elderly. The conversational skills of the domestic robots are currently not comparable to the human level. Only a few researches have been carried out in order to study how the decisions regarding the conversation are made. Humans use a wide variety of non-verbal entities to determine the state of the conversation and manage the conversation flow to ensure that it is interesting to all the parties involved.

The current conversation systems used in service robots frequently use the linguistic message of the user. Most of these systems consider factual and behavior related conversations. Such systems are unable to comprehend the dynamic situations such as the emotional state of the user when generating a response. Systems which can handle the dynamic states of a conversation are preferred by the users. This research is aimed to close the gap between the use of emotions for conversation systems compared

This research proposes a system can be used to make interaction decisions regarding conversation flow by considering the emotional state of the user. In a human conversation both linguistic and para-linguistic features are used to convey information. The proposed system also use this information from the users' response to identify the emotional state and the actual linguistic message. The system comprises of three segments which are emotion recognition, emotional memory storage and conversation decision unit. These three systems coordinate together to decide whether the decision should be continued or not , if it is going to get continued what should be the response and what are the important elements in the user's response that should be retained in system's long term memory.

***Keywords*-Conversation, para-linguistic, linguistic, emotions, service robot, human-robot interaction**

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Dr. A.G.B.P. Jayasekara who provided continuous support throughout the period to successfully finish this thesis. Further I would like to acknowledge my review panelists Dr. Chandima Pathirana and Dr. Ruwan Gopura whose comments and inputs were valuable for my research.

I would also like to thank the members of the thesis review committee who gave their suggestions to improve this thesis.

This research would not have been successful if not for the many suggestions and valuable inputs from my research collegues Chapa Sirithunga, Arjuna Srimal and Viraj Muthugala.

Finally I would like to thank my mother and my sister, both whom assisted in numerous way throughout my research providing me the best environment possible for my research.

This work was supported by University of Moratuwa Senate Research Grant SRC/CAP/17/03.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

The rapid growth of the aging population in the recent years has posed a significant challenge to elderly care sector [1]. As illustrated in Fig.1.1 the elderly proportion of the population is on the rise all over the world. With the professionals involved in elderly care not increasing at the rate of population growth there will be a shortage of caretakers. Further the large number of elderly population will face isolation due to inadequate caretakers. Isolation among elders can lead to deterioration of mental health which can result in depression. Intelligent domestic service robots can be used in elderly care as assistants to care taking professionals or to support the elders for independent living in their own home environment [2]. Both physical and cognitive support for the elderly can be given using domestic service robots with socially communicative capabilities [3].

The intelligent robots that interact with the humans should possess the ability to communicate in a more human-like manner to facilitate smooth bidirectional communication. The service robots are considered as autonomous agents which are designed to carry out complex and unstructured tasks while following the social norms of humans [4]. In recent times several researches have been carried out to integrate a vocal interaction system for domestic robots to enhance human-robot interaction by increasing human likeness. One of the characteristics which is expected from the conversation system includes affective interaction [5].

Most of the existing conversation systems are centered around conversing about factual or behavior related information which do not need adequate cognitive

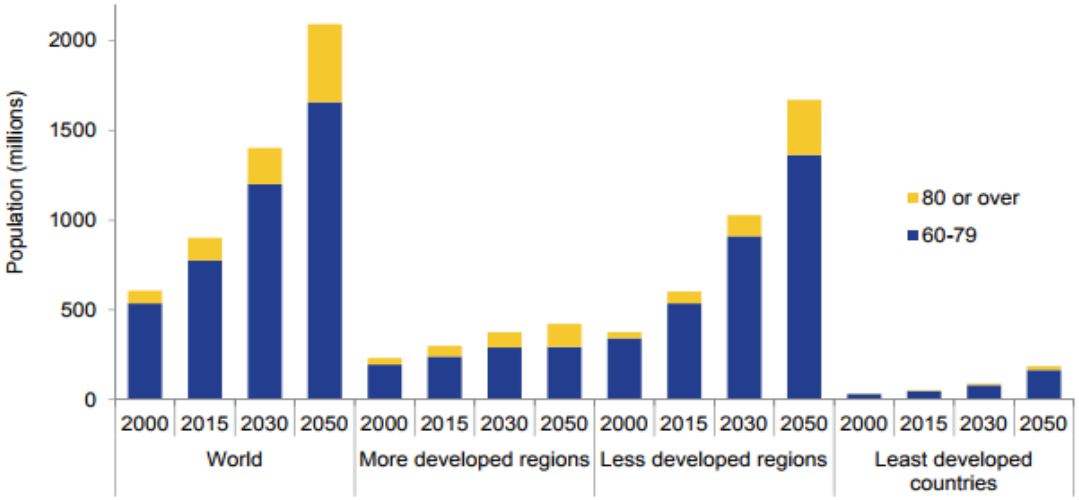

Figure 1.1: Elderly Population Growth Projection until 2050

skills. As the role of the robot expands it is important to enrich its speech capabilities to fit for this larger role. The robots in the future will have to formulate their own opinions or have perceptions about entities similar to their user to be able to be considered a compatible partner to the user. Furthermore the robot should be able dynamically analyze the situation and determine the response without executing the pre-programmed behavior.

One of the most complex tasks for a robot is to be able to understand emotions and act accordingly to the situation. Emotional intelligence is a very important aspect for a robot to be able to be a long term companion for a human. With emotions being an important non-verbal method of communication among humans, empathy or understanding the emotional state of others is considered as an important quality to sustain long term relationships among humans. The ability to detect and understand affective states and other social signals of someone with whom the robot is communicating is the core of social and emotional intelligence.

1.1 Problem Statement

The domestic service robots will have to fulfill their dual role as both a tool and a companion. One of the requirements to be a companion is the robot being able to show and understand the emotions of the user. Empathetic behavior starts from identifying the user's emotional state and adapting its behavior accordingly. The robot behavior can be displayed in many different instances in the daily routine. Emotions do not have a standard definition which makes robotic system to understand what can be perceived as emotion. It is even more difficult for a robotic system to change its action according to the emotional state. Further robots do not have any basis for changing the vocal responses.

Emotions of the user can serve as a form of feedback for the robot it self to modify the behavior. Therefore a robotic system should be able to differentiate between an emotionally significant event and a regular event when communicating with the user.

1.2 Thesis Overview

The thesis is divided in to 9 chapters.The structure of the thesis along with a small description of the chapter is given below.

- Chapter 2 - Literature Review presenting the different past related work of the related areas
- Chapter 3 - Presents the overview of the overall system and its sub systems
- Chapter 4- Presents the vocal feature based emotion recognition system
- Chapter 5- Presents the lexical analysis system which is used to determine the linguistic message conveyed by the user

- Chapter 6- Presents the memory module which is used to store emotional memory elements.
- Chapter 7- Presents how the decisions regarding the flow of the conversation are made.
- Chapter 8 - Presents how the vocal responses are generated in the unit according to the user utterance and emotional state.
- Chapter 9- Conclusions of the research

Chapter 2

LITERATURE REVIEW

2.1 Assistive Service Robots

A service robot can be defined as a robot that performs useful tasks for human or equipment excluding industrial automation applications. Personal service robot can be defined as a service robot for personal use, performing a non-commercial task, usually by lay persons. The common examples for this type include domestic servant robot, personal mobility assist robot and pet exercising robot. The definitions by various sources for the service robots are ambiguous and hence producing a single precise definition is difficult. Yet there are some characteristics of service robots that most parties would agree on. The service robots are expected to have a high degree of autonomy, the ability to navigate through unstructured environments, the ability to perform complex tasks which are not well defined and to interact naturally with humans [6]. Unlike the industrial robots, service robots are expected to understand complex human actions, follow human social norms and to mimic human abilities [6].

Service robot industry is gaining momentum as of recent years. In 2015 service robots for personal and domestic use have a sales value of US$ 2.2 billion which is a 4% increase from the sales value of 2014. The total number of units service robots for personal and domestic use sold was about 5.4 million which is an increase of 16% more than 2014 [3]. It is fore casted that a total of around 42 million units of service robots for personal and domestic use will be sold between the period

Table 2.1: Common Needs and Deficiencies of the Elderly Community

Needs	**Deficiencies**
Guidance	Failing Memory, Disorientation
Physical Support	Muscular-skeletal Fraility, Instability
Health Monitoring	Cardiovascular System, Potential Strokes
Medicine or other Scheduling	Poor Memory

of 2016-2019. The service robots for the elderly and handicap assistance are expected to sell about 37,500 units during the 2016-2019 period and is poised to attain a significant growth rate within the next 20 years. Therefore a significant importance is given for improvements in the service robot industry.

With the expansion of the elderly proportion of the population, the importance of assistive service robots for elderly care has increased. The typical difficulties and the needs of the elders are given in Table 2.1. In order to fulfill these needs the health care sector should have adequate caretakers. It is predicted that the demand of the elderly caretakers will exceed the supply. This may provide a lot of stress on the health sector which in turn will degrade its quality.

2.2 Human-Robot Interaction

In the possible future humans and robots will have to co-exist with sharing and cooperating in various tasks to obtain a common goal. Therefore it is of utmost importance that there should be natural ways that should be used by communicate with the robots in the same way that they communicate with other people [7]. One of the most easier and widely used method of communicating with the robots is through the computer. This method does not resemble the natural mode of human-human communication. Hence it will be an inconvenient communication method for the humans to use on daily basis, especially for the elderly. Multi modal communication provides various channels such as voice, vision and gestures to communicate with the robot. This is significant because

Figure 2.1: Asimo produced by Honda

Figure 2.2: Pepper robot produced by Aldebaran

both voice and gestures are critical elements in human-human communication [8]. A multi-modal system will coordinate with the combined natural input modalities such as speech, touch, hand gestures, eye gaze, head and hand movements to respond with the appropriate output [9]. Hence multimodal interaction can be considered as a part of everyday human discourse [10]. Figures 2.1, 2.3 and 2.2 shows some of the robots that demonstrate skills required for social interaction with the humans.

Figure 2.3: Romeo produced by Aldebaran

2.3 Importance of Conversation In Human-Robot Interaction

Both animals and humans possess social qualities and skills. Yet only humans can communicate through language and carry on conversations which one another. Human conversation skills are evolved into such a great extent that it has the ability to use special characteristics of human body. Human conversations contain complex representations including gestures using hands and head, shifts in body postures, movements of eyes and use of pitch and melody of the flexible voices to emphasize and to clarify what is said [13]. Humans also have the ability to decode such complex representations to understand what the other person is attempting to tell. In fact conversation is a primary skill for humans. Humans start engaging in conversation when they are still infants. Therefore conversation is a powerful way for humans to interact with robots. Hence it is important for the domestic service robots which are to be used in elderly care to possess adequate conversation skills to enhance the quality of human-robot interaction. Especially the elderly population have difficulties in communicating through unfamiliar means such as keyboards or computer screens [11]. In case of developing a conversation system for robots ,it should be based on the study of human-human conversation properties. The engagement model for robots collaborating

with humans include three steps [12].

1. Initiating the collaborative interaction with a human
2. Maintaining the interaction through conversation ,gestures and physical activities
3. Disengaging from the collaboration either abruptly or gradually upon completion of the interaction goals.

2.4 Understanding Emotion

Emotions by nature are defined as complex, fuzzy and indeterminate construct. Emotion Classification is required to distinguish one emotional state from another. Emotions can be described either as discrete or with multiple dimensions. The discrete emotional theory focuses on the the set of six basic emotions which were proposed by Ekman [13] as illustrated in Fig 2.4. The significance of this set of emotions is that they are recognizable across different human cultures and easily distinguishable by individuals facial expressions. Dimensional theories use one or more dimensions to represent an emotional state,unlike the discrete theory it can be used to represent a wide array of emotions [14–16].

The most commonly used dimensional theory is the Circumplex Model which originally proposed by Russell [15]. Fig.2.5 illustrates the simplified version of circumplex model. This theory defines two dimensions which are arousal and valence. Arousal is a measurement of physical activation energy while valence refers to the polarity of the emotion. Depending on the levels of arousal and valence the emotional state can be found. An extension of the circumplex model is the PAD model which introduces an additional dimension of dominance. The PAD model was able to define even more emotions than its preceding theories. More recent theories such as Löveheim's Cube provided a more understanding on

the biological facet of emotions [17].

In order to recognize affect, the first step would be to define what is perceived as emotions. The word emotion has hundreds of definitions which were proposed throughout the past few decades. Further affect and emotion are both used interchangeably in most of the literature. Affect can be considered as the predecessor to emotions and feelings [18]. Affect is a non-conscious experience that determine the relationship between the person, environment and people [19]. Affect can further be described as a positive or negative assessment of an object, behavior or an idea [20]. Affect is considered difficult to detect since it is an internal state. Hence all the affect detection systems are usually designed to detect emotions.

Emotions can be considered as preconscious expressions of affect that is influenced by culture [18]. Dolan [21] presented three characteristics of emotions, being full body experiences, harder to control and having a global impact on the person. Plutchick stated that emotion as a complex chain of events which begins with a stimulus [22].

Most of the current emotion detection systems uses one or several modalities. There can be several factors which contribute to the choice of a specific modality such as the validity of the signal as a natural way to identify the emotion, reliability of the signal in real world environment, the time resolution of the signal and cost for the user [23]. Facial expressions can be considered as the most widely used emotion detection method referred in literature. Further the six basic emotions as stated by Ekman [24] are universal and can be recognized through cross cultures easily with the facial expressions. Additionally other modalities used to recognize affect include speech, body language, posture, physiology, brain imaging and sometimes a combination of several modalities [23] [25] [26].

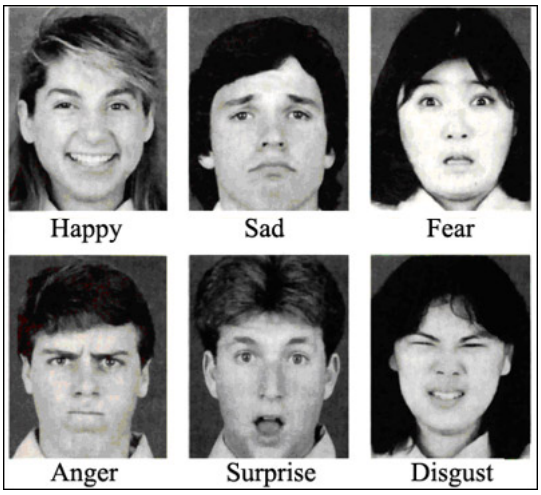

Figure 2.4: Six basic emotions presented by Ekman expressed through facial expressions

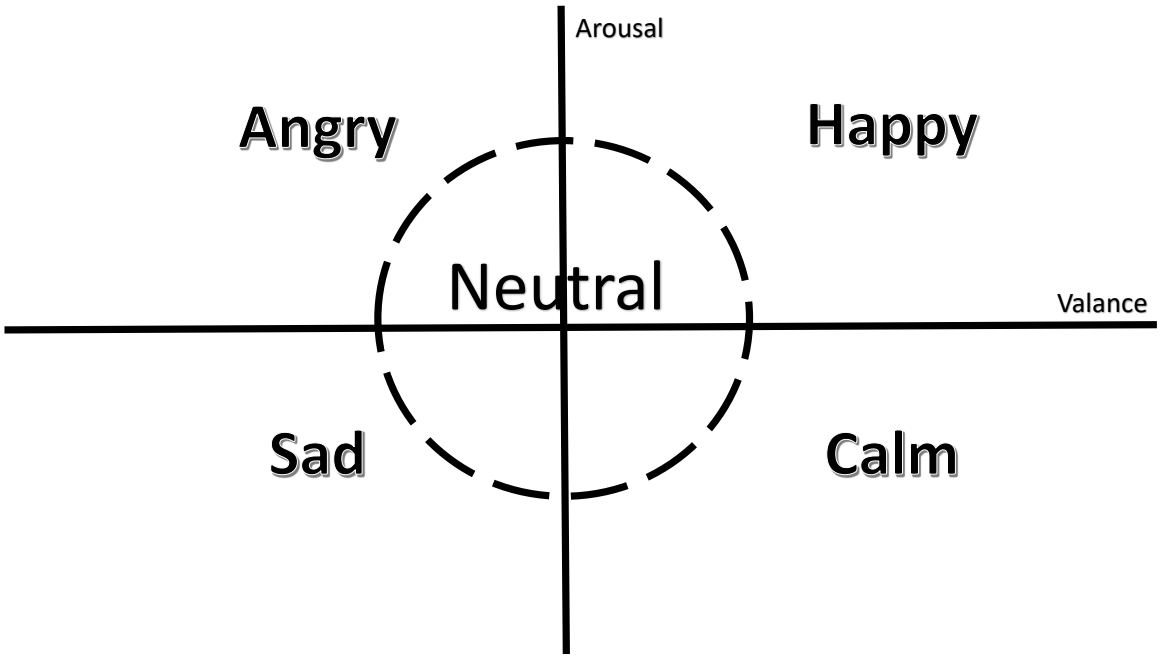

Figure 2.5: A simplified version of the Circumplex model introduced by Russell

2.4.1 Emotion Classification Systems

Speech based Systems

Due to its low cost and non-intrusive nature emotion recognition using speech can be considered promising [23]. Speech based emotion recognition system can consist of two parts as the feature extraction unit and the classification unit. Emotions can be conveyed through speech both by the linguistic message and the implicit paralinguistic features. Most of the systems are based on the paralinguistic features since extracting the linguistic messages require additional processing. Speech features can be divided into four categories which are continuous features, qualitative features, spectral features and Taeger energy based features [27]. The optimal set of vocal features for emotion recognition is not yet determined. However there are specific vocal features for certain emotions that makes them easier to distinguish among others. Further it is important to note that states such as boredom and frustration can be identifiable from nonlinguistic aspects such as sighs and yawns [28].

The features extracted from the speech signals can be either global or local. Global features are calculated as the statistics of all speech features extracted from the utterance [27]. Local features are calculated after creating segments also known as frames of the speech signals and extracting the features from these segments. Prosody continuous speech features such as pitch and energy as well as voice quality features are strongly related to emotional content [27] [29] [30]. Prosody and voice quality features are preferred for emotion recognition in simulated instances while spectral features are used in natural speech [31].

Classification schemes frequently used for speech emotion recognition includes Hidden Markov Models(HMM), Gaussian Mixture Models(GMM), State Vector Models(SVM) and Artificial Neural Networks (ANN). In speech based emotion recognition systems there are no preferred classifiers and the choice of classifi-

cation scheme depend mostly on the nature of the extracted features and the sample size. Among these classifiers ANNs are usually preferred over HMMs and GNNs when the training size is relatively lower [27]. GMMs are used for speech emotion recognition when global features are used.Among the ANNs, Multi Layer Perceptron(MLP) and Recurrent Neural Networks(RNN) are the types used for speech based emotion recognition. Further use of several ANNs and then combining them using an appropriate aggregation scheme to recognize emotion can also be found [32].

Other classification schemes used to a lesser extent are fuzzy classifiers [33] [34] and k-NN classifiers [35] [36]. A more computationally intensive and complicated method will be to combine several classifiers instead of a single classifier to increase the accuracy [37]. The classifiers can be connected in a hierarchical, serial or a parallel manner [27].

Robotic systems such as Kismet were able to recognize affective states through voice and also display certain emotions [38].

Vision based Systems

A large proportion of the affect detection systems use facial expressions to detect the basic emotions. In order to use facial expressions it should be assumed that each emotion will display a distinctive facial expression. Detecting emotion through facial expressions is a matter of detecting the prototypical facial expression when the emotion is triggered [23]. All most all the vision based emotion recognition systems are focused on identifying the six basic emotions.

Majority of the facial features used in emotion recognition systems are either geometric features such as shape of the facial components or appearance features representing the facial texture such as wrinkles and furrows [39]. However systems which use both geometric and appearance features have also being proposed which are considered better option for emotion recognition [40]. A variety of classifiers

including SVM, HMM, ANN and sometimes a combination of several of them.

Text based Systems

Text based emotion and sentiment analysis has strong roots in social engineering and such algorithms are widely used in social media to analyze the perception of the users. However in human-robot interaction the text refers to the written transcriptions of the conversations between the human and robot. The research on text based affect detection started early with multidimensional scaling (MDS) to create visualizations of affective words based on similarity ratings [41] [42].

Lexical analysis of text is another prominent method that is used to identify words that can convey emotional states of the writers or the speakers [43]. Sentiment analysis or opinion extraction is another widely used complex methodology which focuses on the valence of a text without assigning a particular emotion to it. Sentiment analysis is done by building a affective models from a large word corpora and apply these to find the tone of the text.

Hybrid Systems

Hybrid systems use data from more than one modality and performs a fusion mechanism. Data fusion can be done either at the feature level or decision level [44]. In the feature level sets of features extracted from each sensor are combined together. When determining the feature set it is important to include the unique features of each sensor [23]. In the decision level fusion is done by merging the classifier output of each sensor [45]. Hybrid systems usually have higher accuracies compared to systems using single modality and is a much better emulation of the human emotion recognition system.

Bhasker et al. presented a hybrid system based on State Vector Machine (SVM) classifier using vocal features and text transcripts of conversations to

Table 2.2: Stimulus events and Expected Behavior of Humans

Emotion	**Stimulus Event**	**Expected behavior**
Happy	Gaining of a valuable object, Reminiscing a delightful memory	Sharing the happiness
Angry	An unpleasant event, An objection	Managing Anger
Fear	A threat	Eliminating threat
Disgust	An unpleasant object, event or a person	Eliminate the source of disgust
Sad	Loss of a valuable person or an object	Consoling the person
Surprise	An unexpected event	Expressing surprise

classify six emotional states [46]. Further Schuller et. al. also proposed a hybrid system which uses a SVM and Belief Network based architecture for emotion classification using acoustic and linguistic information [36]. Considering bimodal systems using both visual and audio channels Busso et.al. [47] proposed a such with fusing data in both feature level and decision level.

2.5 Verbal Response Generation According to the Emotions

Emotions can naturally serve as a form of feedback for a robot when behaving with the humans. For humans the emotions convey an underlying opinion about the service or the opinion express. All emotions are expressed due to a simulating event [14]. Table 2.2 shows the possible stimulus events and expected behavior for the emotions.

When the robot is interacting it should respond to the stimulus effect of the emotions. Understanding the stimulus effect of the human can have a wide scope which refers to the discourse of the discussion. During the discourse the human might have expressed the emotional stimulus. Therefore tracking the discourse of the conversation is important in understanding what event caused the emotion.

The conversational agents can be divided into two categories which are goal

based dialog agents and chatbots. The goal based dialogs are built mainly using two architectures which are

- Finite State - In this kind of dialog management the system completely controls the conversation with the user. Such systems are also known as system initiative or single initiative systems. These systems ask the user a series of questions while ignoring anything the user says that is not a direct answer to the system's questions. Such systems are easier to develop since the user knows what they can say and the system knows what the user can say. The functionality of such systems are limited.

- Frame Based - These dialog systems are mixed initiative in which the initiative can pass between the user and the agent. In such systems the user can answer multiple questions at the same time unlike finite state systems. These systems are an improvement from the finite state systems as they avoid strict constraints on order of the conversation flow.

Chatbot architecture can be either rule based or corpus based. The rule based systems mostly use a pattern-action rules such as ELIZA [48] or a mental modal such as Parry [49].

The corpus based methods can be further divided into information retrieval and deep learning modals.

- Information Retrieval- These systems mine conversations of human-human or human-robot chats. The system's objective is to find a prior turn that matches the user's turn and give the corresponding response.

- Deep learning chatbots- These systems train on twitter or movie dialogue databases. Sequence to Sequence model architecture is the most widely used for such chatbots [50].

2.6 Current status on existing human- robot vocal interaction system

Due to technological limitations the conversations between humans and robots are simpler and more constrained than human-human conversations. Yet robots with a limited conversational capabilities have been used as supermarket assistants [51], receptionists [52], elderly caretakers [11] and gym instructors [53]. These robots engage in conversations associated with giving directions, giving factual information, engage in small talk and performing memory games [11] [51] [52]. It is noted that in all the above instances the conversations were task oriented .

Most of these robots used a goal based or frame based approach in conversation rather than using any other method. Hence the replies which the robot presented were standard. Further there is a general disregard among the robots in conversation about the emotional state of the user. Therefore by not being sensitive to emotions the robots will not be able to share a relationship with a human as a companion. Although robots who are able to display emotional expressions through facial or any other physical features have been designed, emotions being conveyed through non-preprogrammed conversations is almost scarce.

2.7 Summary

The service robots which have to be used in elderly care should be used as long term companion. The long term companion robots have to adapt according to the personality traits of the user to co-exist by building the robots own personality. Further in order to keep the conversations fresh and interesting the robot should have a mechanism to update the knowledge base giving priority to the interested areas of the user. When carrying on the conversation the robot have to monitor the interest level of the users by using facial expressions or gestures and change the topic to a more interesting area for the user. The conversational domestic

service robots are mostly beneficial for the elderly people. These conversation systems should have the ability to change the flow of the conversation according to the user's needs which is done with little flexibility in current systems. Overall in most cases the emotions are disregarded during the conversation flow or else with limitations. Table 2.3 presents the summary of the literature review

Table 2.3: Summary of the Literature Review

Area	**Limitations of the current system**	**Possible Improvements**
Scope	Limited Response generation capability without considering the emotional content	Use an emotion recognition method when generating responses and making conversation decisions
Communication	Uses either linguistic or emotional state of the user	Integrating linguistic and para-linguistic features for better understanding of the user's state
System Adaption	Adapt the behavior accordingly with experience	Use a neural network based methodology which can improve itself by using the data from past conversations

Chapter 3

SYSTEM DESIGN

3.1 Studies of Human Conversation

Modeling human-human conversation is an extremely difficult task due to its intricate and complex nature. Conversation is a joint activity between two or more interlocutors. The conversations are gradually built up by consecutive turns. Each of these turns consists joint action between the speaker and the listener. The listener make inferences called as conversational implicatures about the intended meaning [54].

Although it is difficult to identify all the properties required for a proper dialogue system the following properties can be considered important in human-human conversation [55].

- Understanding both verbal and nonverbal inputs
- Generating verbal and nonverbal outputs
- Dealing with conversational functions such as turn-taking , conversation fillers and repair mechanisms.
- Giving signals that indicate the state of the conversation
- Contributing new suggestion for the discourse.

If the robots are to be long term companions they should additionally possess

the abilities such as being sensitive to affect, evaluate consequences of actions, justifying decisions and adapt to the individual [56].

3.2 An Artificial Agent Conversation System

A simplified dialogue system which performs a task oriented conversation which is widely used in commercial applications, typically contain six components. Fig.3.1 illustrates the basic structure of a conversational agent which uses only the user's linguistic response as the input.

Speech Recognition - This component takes an audio input generally from a microphone and returns a transcribed string of words as the output . In a case where automatic speech recognition module output has an unacceptable accuracy as in a noisy environment Wizard of Oz method can be used. The sentences that the speech recognizer needs to be able to transcribe are just those can be understood by the Natural Language Understanding Component.

Natural Language Understanding(NLU)-The NLU component of dialogue system must produce a semantic representation which is appropriate for the dialogue task.

Dialogue Manager- The dialogue manager system controls the architecture and structure of the dialogue. The dialogue manager takes the input from the NLU unit, keeps track of the state of the conversation, interfaces with the task manager and passes the output to the Natural Language Generation (NLG) Unit. The final output of the dialogue manager is to decide what to say. In some cases a seperate content manager module is merged with the dialogue manager decides on what content to express to the user, whether to ask a question, present an answer, and so on.

Task Manager- This component holds the knowledge base associated with the

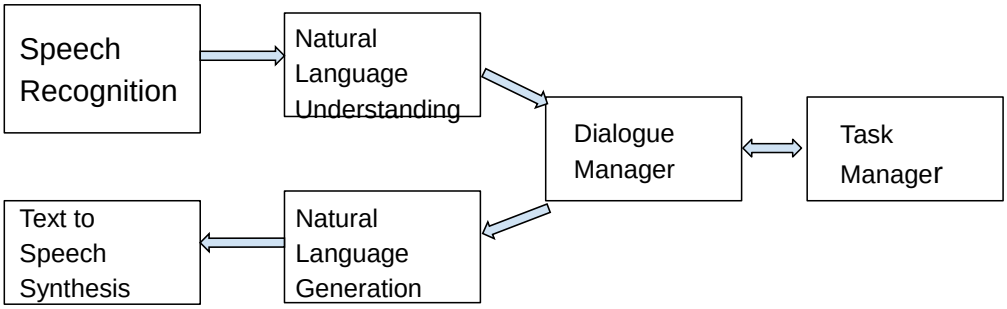

Figure 3.1: Simplified architecture of a dialog system

task the for which the system is designed.

Natural Language Generation (NLG)- The NLG unit has to decide on how to present the content the dialogue manager has passed to the NLG unit. This unit chooses the syntactic structures and words to express the meaning.

Text to speech Synthesis-This component then takes these words and their prosodic annotations and synthesizes a waveform.

This basic system does not imitate the human-human conversation and has severe constraints like strict grammar rules,lack of regard for emotions, disregarding the knowledge gained from the previous conversations.

3.3 System Overview

The proposed conversation system in this research integrates an audio based emotion recognition unit along with a conversational unit. This integration will enable the system to use the speech modality, which enables the system to use para-linguistic features when making decisions regarding the conversation. The information retrieved from this unit is then taken into consideration when the decisions regarding the conversation are made. The conversation system contains

five subsystems which are given below

1. Audio Based Emotion Recognition
2. Linguistic Analyzer
3. Memory Storage
4. Interaction Decision Unit
5. Vocal Response Generator

These subsystems are discussed in detail in the following chapters 4,5,6,7 and 8. This system uses only the audio clips acquired through a blue tooth headset and microphone combination. By using the audio stream only the system can function much faster than the video stream. Figure 3.3 presents the overall system with the integrated subsystems.

This system will not only be able to determine what is said but also to determine how it is said to have a better understanding of the user's emotional state. This will enable the conversation system to make decision's in a more personalized manner towards the user.

The conversation system was built with a graphical user interface for easier handling. The software module is named MIVoIS (Moratuwa Intelligent Vocal Interaction System). The module was developed in a Linux environment and written entirely in Python. The MIVoIS module has the ability to run tests on individual units as well as the system as a whole. The MIVoIS module is mainly used for testing the system with the user as it provides a clean interface. Figure 3.2 shows the home screen of the module. This module also gives a much smoother control and less complicated front end for the conversation system.

The MIVoIS program currently has three sub modules. All these modules are used to test individual units such as well as to record the interaction for future training and improvements.

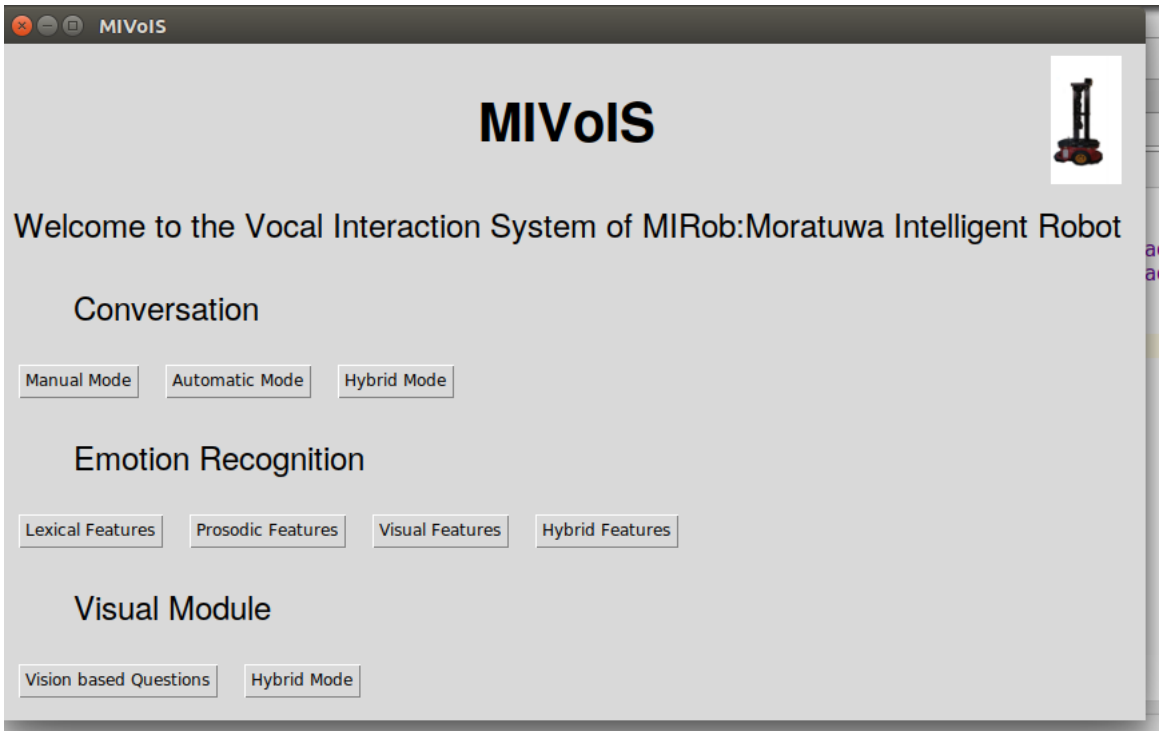

Figure 3.2: Home Screen of the MIVoIS conversation module

- Conversation Module : This unit can is used to test the overall system output. All these modules have the ability to save the conversations in transcript form which is used for training the system to improve the performance overtime.
  - Manual Mode: In this mode the operator has the total control over what the robot says and is frequently used to obtain test data from practical scenarios.
  - Automatic Mode: In this mode the operator does not have the control over what the robot says and whatever the response produced by the system is taken as the response.
  - Hybrid Mode : This module gives the option for the operator to whether to use the system generated utterance or any appropriate utterance for the situation. This module was used to gather test data for training.

- Emotion Recognition Module :
  - Lexical Features : This module facilitates the Lexical analysis and a sentiment analysis on the linguistic message.
  - Vocal Features : This module is used to test the functionality of the vocal feature based emotion recognition.
  - Visual Feature : This module is for future development to integrate the visual features into the conversation system.
  - Hybrid Features: This module uses both linguistic and vocal features to determine the emotional memories of the user.
- Visual Module: This module is for future developments where the system will be able to converse about what it sees.

3.4 Summary

This section presents a brief overview of the overall system and the included subsystems. The following chapters will describe the individual subsystems in a more detailed manner. This conversation differs from the traditional systems in the sense that it takes the emotional state of the user into consideration when making decisions regarding conversation and producing the responses. This system will be able to be conscious of the user's emotions which will be beneficial in developing the emotional intelligence of the robot. The MIVoIS module provides a user friendly front end for the conversation system which can be used for both system and unit testing. The following chapters will provide a more detailed view into each of these subsystems.

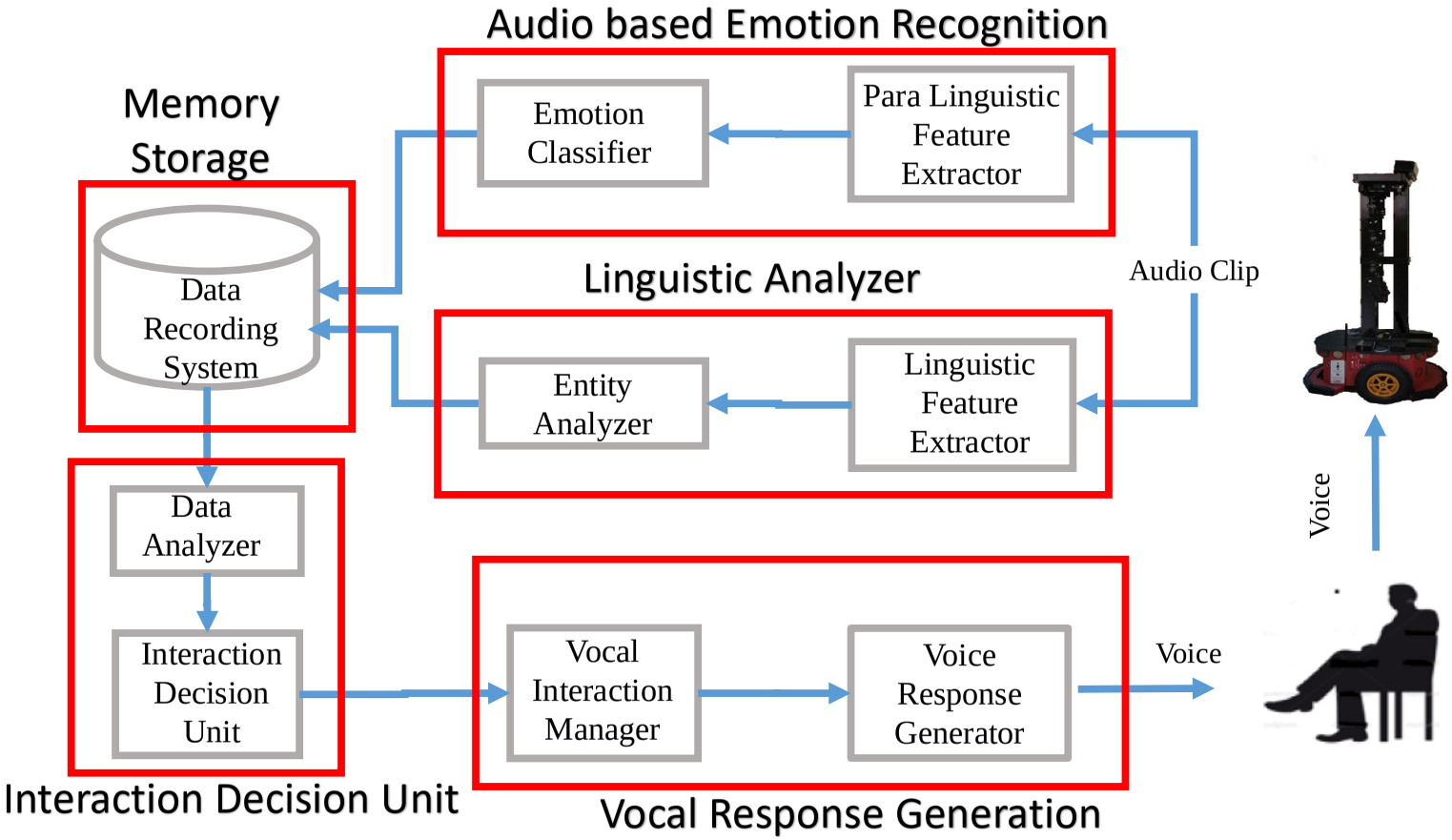

Figure 3.3: The overall conversation decision System

Chapter 4

EMOTION RECOGNITION SYSTEM

Speech is a very important mode of communication which is used frequently used by humans as their primary method of interaction. Speech can transmit information through two methods.

- Explicit linguistic message - *what is said*
- Implicit paralinguistic message- *how it is said*

This chapter focuses on the para-linguistic features of speech which can be primarily used to decode emotional content of speech. Para-linguistic features that can be used for optimal emotional recognition are not yet fully determined. Further there is a disagreement among the researchers about which acoustic features are more suitable in emotions. Further this chapter includes emotion recognition systems using both prosodic and spectral features.

This chapter provides answers for the following questions:

- What emotions can be detected ?
- What features should be used for Emotion Detection?

4.1 Selection of Emotion Space

As the first step of the system design the emotion space must be defined. As only the voice component is only used all the emotions defined in literature cannot be displayed and represented in the system. A participant can show a range of emotions when communicating which as shown in literature can be more than 100 categories. Humans use a combination of speech and other physical features in order to convey emotions. Therefore it is not possible to represent all the emotions with using vocal characteristics of a person.

As the first step it is important to determine the emotional states that can be distinguishable by vocal features of the user's vocal response. For this experiment 1721 audio samples of 7 emotions which belonged to anger,happy,sad, disgust, fear, surprise and neutral were used. These emotional states were chosen to represent the 4 quadrants of the Circumplex Model mentioned in 2 so that the whole emotional space can be covered. Emotions such as disgust, fear and surprise were do not have a specific position in the emotion space.

4.2 Linguistic vs Para-Linguistic Features

The first question that is encountered is to decide on linguistic or para-linguistic features of a speech as this is the first categorization.The advantage of using a para-linguistic feature based method over a lexical feature based method is that the users can use the same utterance in different affective states. Hence affective classification cannot be done purely based on lexical features. As Fig. 4.1 shows the pitch and intensity contours of 4 audio samples that were obtained during the experiment. These represents four affective state even though the participant uttered the same words.

Therefore emotion classification based on lexical features cannot be considered

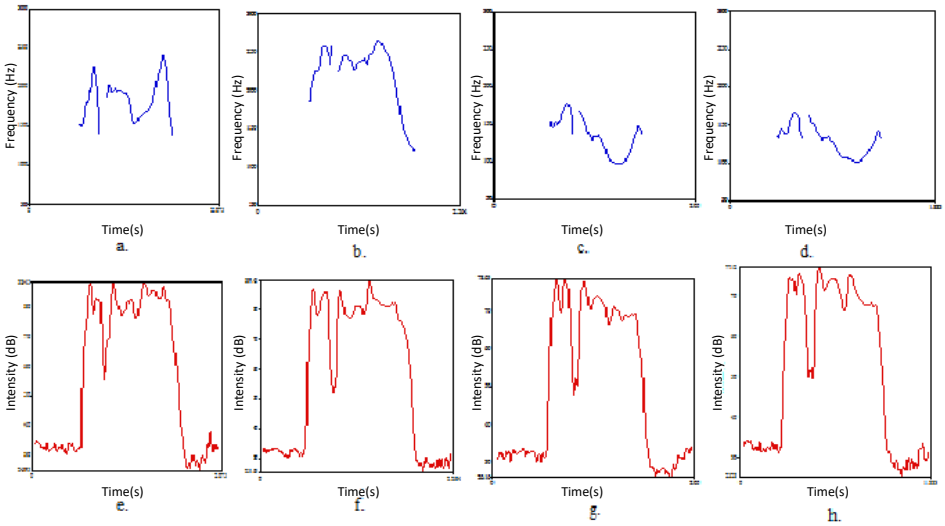

Figure 4.1: Pitch contours(upper row) and Intensity Contours(lower row) for the four affective states when the user presents the same utterance a) & e) Angry b) & f) Happy c) & g) Sad d) & h) Neutral

suitable for such system. Further the lexical features can only give us an indication of the valence of positivity or negativity of the statement. It cannot be used to determine the affective states which differ in energy levels. Therefore it was decided that para-linguistic features should be used for the emotion recognition as both valence and arousal levels should be considered for emotion recognition.

4.2.1 Feature Extraction for Emotion Recognition

The features used for the emotion recognition can belong to either spectral or prosodic features. The input for the classification system is the audio clip of the user's response. The audio-clip is then fed separately to the prosodic feature and lexical feature extracting units. This unit focuses only on the feature extracting methodology and the associated emotional classifier.

Figure 4.3 illustrates the layered structure of the system that is associated with the audio system. In the bottommost layer is the pre-processing unit which apply filters to eliminate unwanted noise and disturbance. The pre-processing is done optimally and with great care to ensure that the filters do not eliminate the

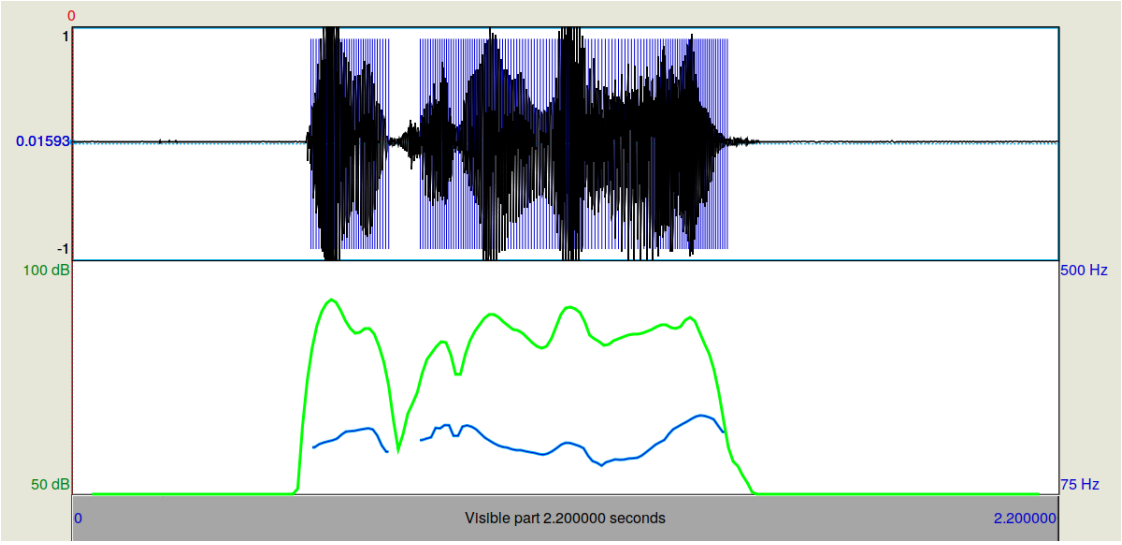

Figure 4.2: Visible voice, Pitch and Intensity Contour for an Angry Utterance

emotional content of the audio responses.

4.2.2 Extraction of Prosodic Features

Prosody is associated with larger units of speech or sequence of speech sounds. Prosody can be used to understand the features related to the speaker such as emotional state or related to the utterance such as the form. Prosody is important in recognizing emotions that can be linguistically difficult to distinguish.

Prosodic parameters can be analyzed by using either auditory or acoustic measures. The auditory measures are associated with the impressions produced by the listener's mind and can be measured using auditory scales such as mel scale. The acoustic measures can be obtained through the physical properties of the sound wave. In this model only acoustic parameters are considered which are given below .

- Maximum Pitch (Hz)
- Minimum Pitch (Hz)

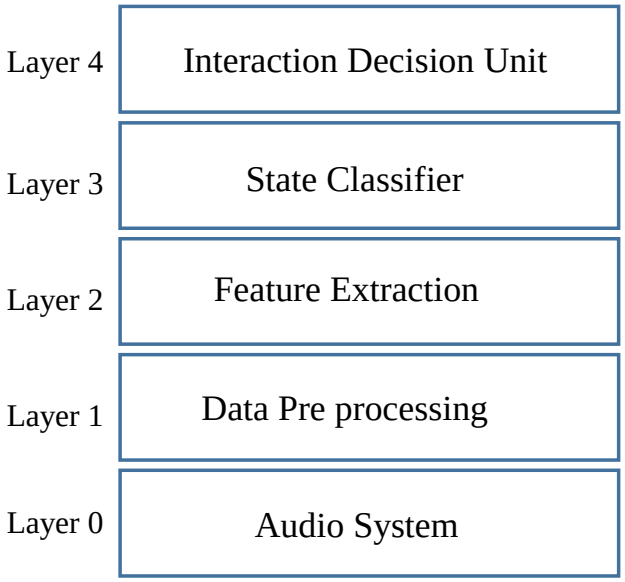

Figure 4.3: Layered Structure of the System

- Mean Pitch (Hz)
- Speech Rate (syllables per minute)
- Amplitude Root Mean Square (Pa)
- Mean Power (dB)

4.2.3 Extraction of Spectral Features

There is a certain doubt in previously conducted research that whether local or global features of audio clips. The majority of researchers agree upon the global features being superior than local features regarding classification accuracy and time [27]. Global features can be considered as a representation of statistics regarding the total vocal response. Features were chosen in order to represent the emotional content of the audio waveforms. The features were extracted using Pyaudio and PyAudioAnalysis library [57].

The audio clips are typically 3 to 7 seconds long. The audio clips are then fractioned into fixed length 100ms segments. Some segments specifically the fi-

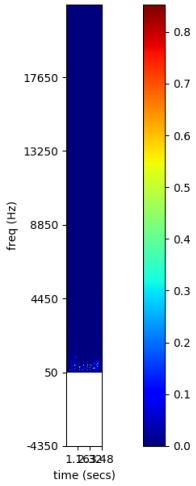

Figure 4.4: Spectrogram of the System

nal segment of most of the audio clips are less than 100ms. All the utterances that were considered were simple sentences. Further the entire utterance corresponds to a single emotion and no utterances were used which include switching of emotions within the sentence.

For this research following features were extracted from the audio clips.

- Zero Crossing Rate- The rate of sign-changes of the signal during the duration of a particular frame.
- Energy-The sum of squares of the signal values, normalized by the respective frame length.
- Entropy of Energy-The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
- Spectral Centroid-The center of gravity of the spectrum.
- Spectral Spread-The second central moment of the spectrum.
- Spectral Entropy-Entropy of the normalized spectral energies for a set of sub-frames.

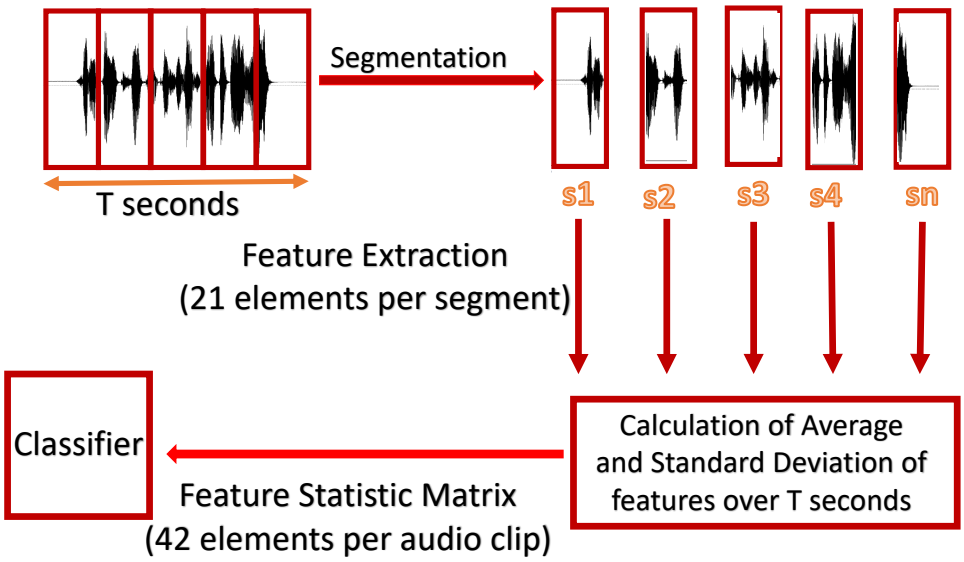

Figure 4.5: Feature Extraction Methodology

- Spectral Flux-The squared difference between the normalized magnitudes of the spectra of the two successive frames.
- Spectral Rolloff-The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
- Mel Frequency Cepstral Coefficients (MFCCs)- This is a representation of the short term power spectrum of a sound. There will be 12 MFCC values considered.

These features were extracted after segmenting the audio clip and obtain the average and standard deviation of each of the features over the total time of the audio clips duration. Figure 4.5 illustrates how the system prepares the feature matrix.

4.2.4 Identifying Emotion Clusters- Using Unsupervised Learning

As mentioned in the literature survey all the emotional states cannot be correctly identified using the vocal parameters. Further all the test data cannot

be sometimes labled.Hence as the first step of Fig.4.6 illustrates the clustering process done using unlabeled data to determine the emotional states that can be determined using using vocal features. The clustering process is done in several stages as shown in Fig.4.6.

- **Stage 1:** *Feature Extraction* The prosodic and spectral features for each audio clip was extracted separately and combined to prepare the emotion feature input file. Several data files were used in the process with data from 7 to 4 distinctive emotions.

- **Stage 2:** *Normalization* The features in the input file are numerical values in different scale, hence data normalization is performed in order to bring all the values to a common scale. In this case normalization was done using z-score method.

- **Stage 3:** *Clustering* In order to determine how many emotion clusters can be formed by the existing data. In this case K means++ clustering algorithm was used. This algorithm was run using the total parameter range since at this point the specific features required in order Eucledean distance was used as the metric which was chosen to measure the distance between the chosen centroids.

- **Stage 4:** *Sweep Clustering* This module is needed to perform a parameter sweep to determine the optimum settings for a clustering model. Both the normalized data set and clustering algorithm are given as inputs. The setting used are as follows-

  Parameter Range Entire Grid of Parameter Range since the optimum features for emotion recognition is not known at the time.

  Metric for measuring Clustering Result Four methodologies were used including Simplified Silhouette,Davies-Bouldin,Dunn and Average Deviation Method on separate occasions to optimize the result. It is also

noted that the results did not show a significant difference among the 4 methods.

The output of this module are the distances between each point and the centroids of the cluster.

- **Stage 5:** *Assign Data to Clusters* This module will take the output of the sweep clustering module and make the assignments according to the distances.

- **Stage 6:** *Applying SQL Transformation* This stage is needed for better visual representation of the clusters.

At the end of the process the principal component analysis (PCA) graph can be created to visualize the dimensionality of the clusters. Fig.4.7 shows a PCA graph generated with the clusters corresponding to the emotions Neutal(Cluster 3), Angry(Cluster 2), Happy(Cluster 1) and Sad(Cluster 0).

- The first component axis is the combined set of features that captures the most variance in the model. It is plotted on the x-axis (Principal Component 1).

- The next component axis represents some combined set of features that is orthogonal to the first component and that adds the next most information to the chart. It is plotted on the y-axis (Principal Component 2).

The PCA ellipses for the clusters are all oriented in different directions, indicating that there is some separation between them though clusters 0 and 1 are not as well separated as clusters 3 and 2.

This was the best performance that could be gained from the clustering system was when only 4 clusters were considered. Therefore it was decided that the emotions Happy, Angry, Neutral and Sad will be used as the emotions.

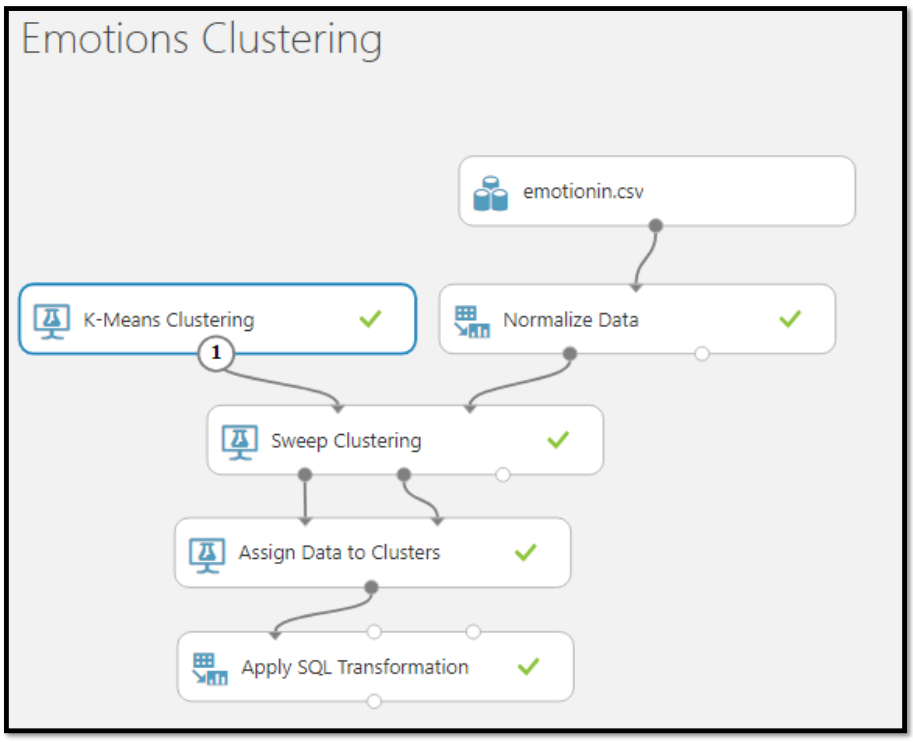

Figure 4.6: Clustering Process To Determine Vocally Distinguishable Emotions

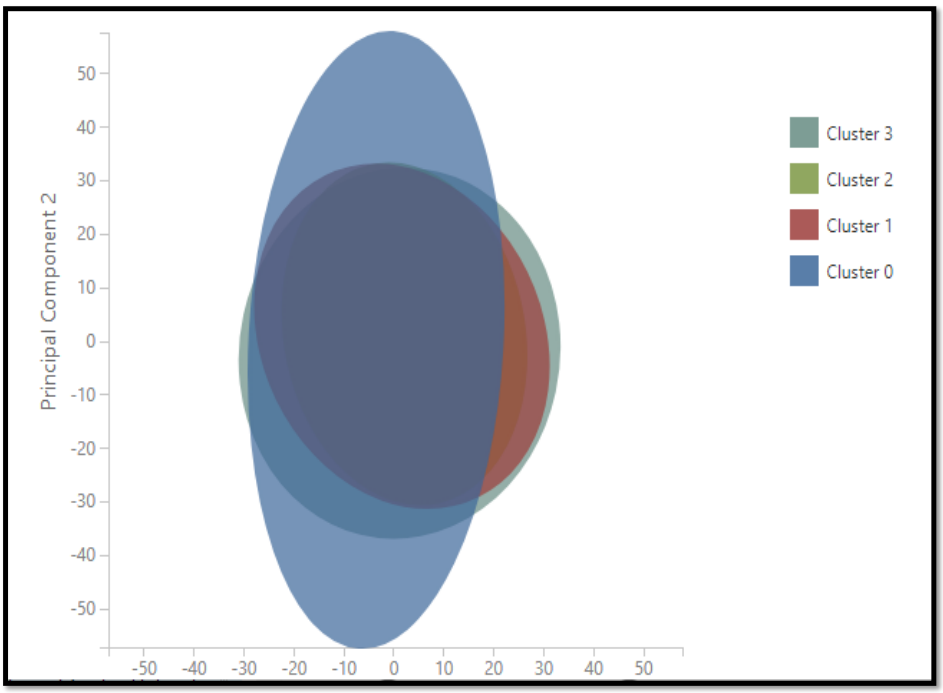

Figure 4.7: PCA (Principal Component Analysis) graph showing 4 clusters

4.3 Neural Network for Audio based Emotion Recognition

In the second phase a neural network was trained after extracting the features mentioned in section 4.2.3 and using them as inputs. In the second phase a neural network There are 20 spectral features and 7 prosodic features to be considered of which making a 47 input neural network.

Each of these neural networks are trained separately so that each of them represent an expert of recognizing that particular emotion.This procedure of the neural network is illustrated in Fig.4.8. For each neural network the weights and variables are initialized to random values and then optimized.

- **Stage 1:***Input File* The prosodic and spectral features for each audio clip was extracted separately to prepare the emotion feature input file. Several data files were used in the process with data of 4 distinctive emotions.
- **Stage 2:** *Principal Component Analysis* With the large number of features associated with the Principal component Analysis is required to reduce the dimensionality of the features. For spectral features the MFCC values were combined to reduce the dimensionality.
- **Stage 3 & 4:** *Normalization* Normalization is done in two parts because the differences of the distribution of features. For normally distributed features Z score method is used and Min-Max method is used for skewed distributions. Thus bringing all the features to the same scale.
- **Stage 4:** *Multiclass Neural Network* This module is needed to perform a parameter sweep to determine the optimum settings for a clustering model. Both the normalized data set and clustering algorithm are given as inputs. The setting used are as follows-

  Parameter Range Entire Grid of Parameter Range since the optimum features for emotion recognition is not known at the time.

Hidden Layers One hidden layer

The output layer is fully connected to the hidden layer.

The hidden layer is fully connected to the input layer.

The number of nodes in the input layer is 29 for spectral features and 7 with prosodic features.

The number of nodes in the hidden layer is set to 25.(This was decided after changing the number of hidden nodes in the range of 15 to 100 and comparing the overall accuracy)

The number of nodes in the output layer is 4 corresponding to the number of emotions going to be detected.

The output of this module are the distances between each point and the centroids of the cluster.

- **Stage 5:** *Data Splitting* This module will take the input data and split randomly , with 70% of the data for model training and the rest for model evaluation.

- **Stage 6:** *Tune Model Hyperparameters* This stage is needed since we do not know what are the features that represent the emotions best. This module will look at the label values and determine the best features which represent the emotional content.

- **Stage 7:** *Score Model* This stage is produces the predictions using the trained model.

- **Stage 6:** *Evaluate Model*The final evaluation with the results are done and the confusion matrix is generated.

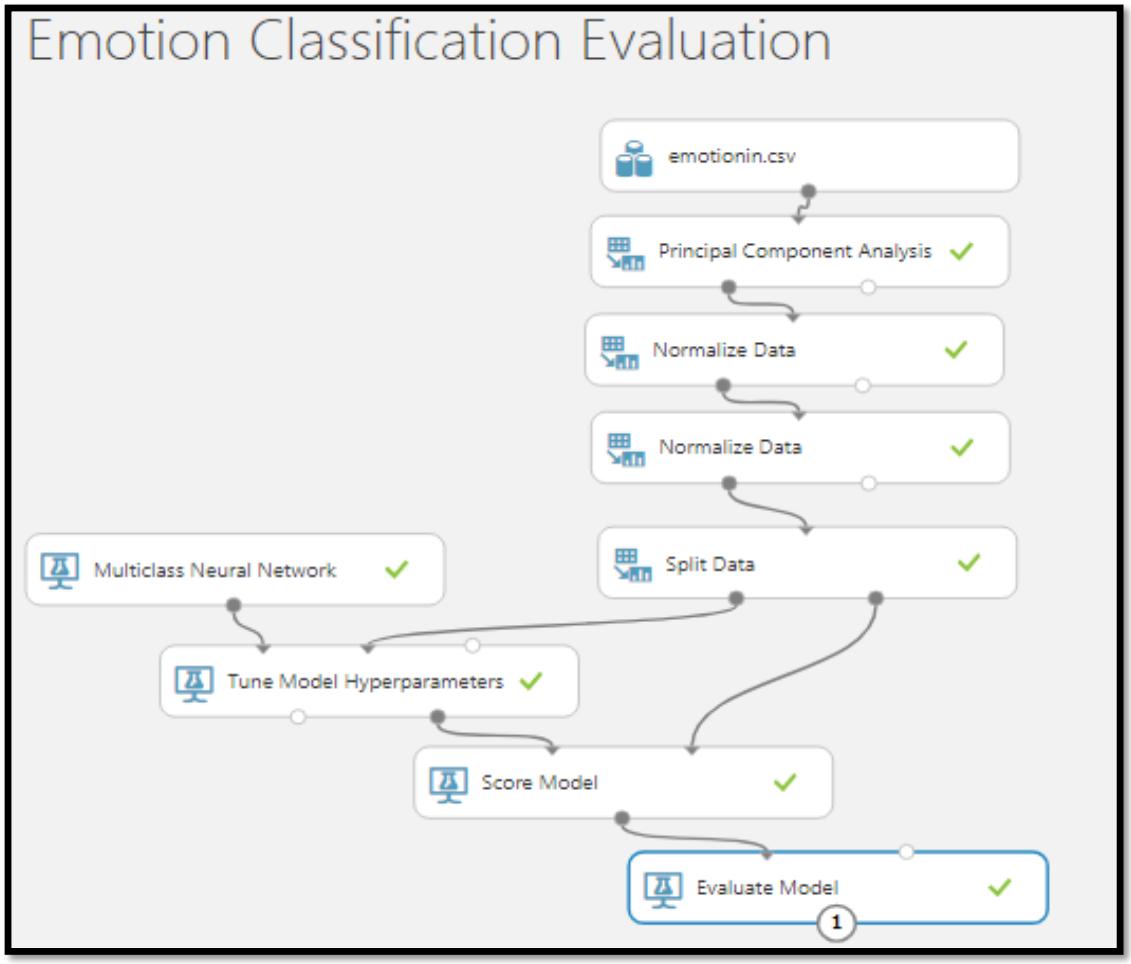

Figure 4.8: Flow of the neural network based Emotion Recognition System

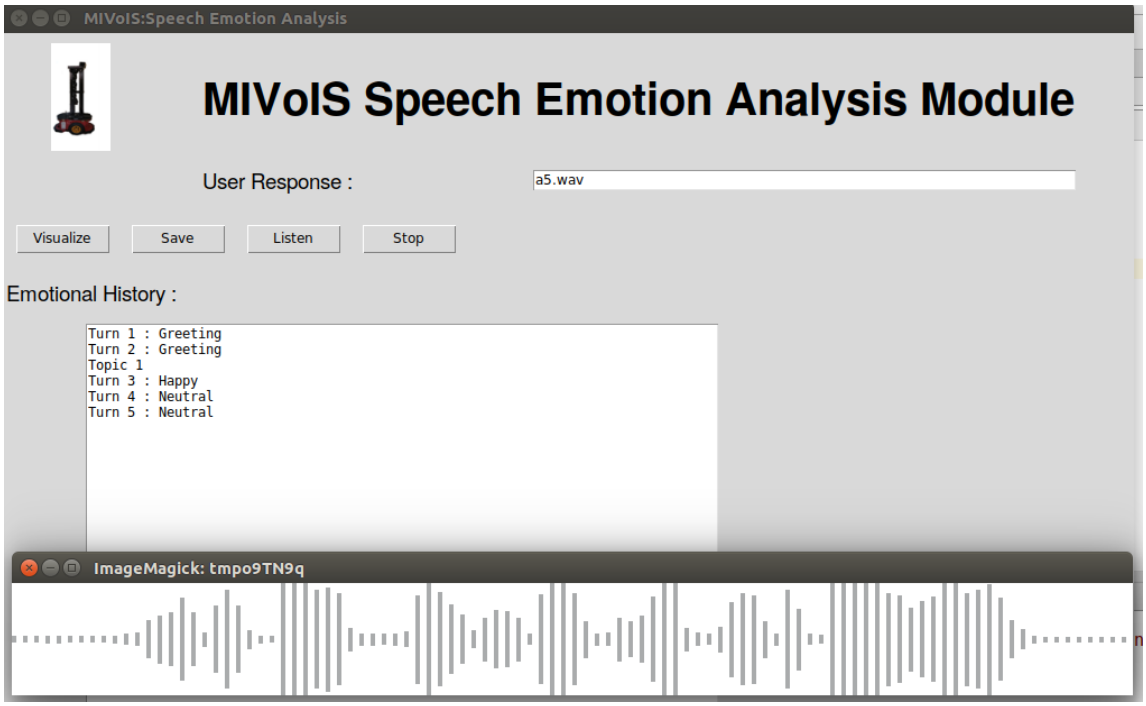

Figure 4.9: Visible voice, Pitch and Intensity Contour for an Angry Utterance

4.4 Implementation of Emotion Recognition System

The trained model is hosted in a cloud service for better compatibility across all the devices. Hence the emotion recognition module is configured to run as a web service. Fig.4.10 shows the back-end of the emotion recognition module. This service will take the numerical values for the neural network input parameters as a vector and the output will be the label of the emotion which will again be delivered through the web service.

The front end of the Emotion recognition system is designed with a graphical user interface to make it easier to work with the system. Figure4.2 shows a screen shot of the graphical user interface of the front-end. The user response is taken as the input in .wav format for feature extraction through this interface.

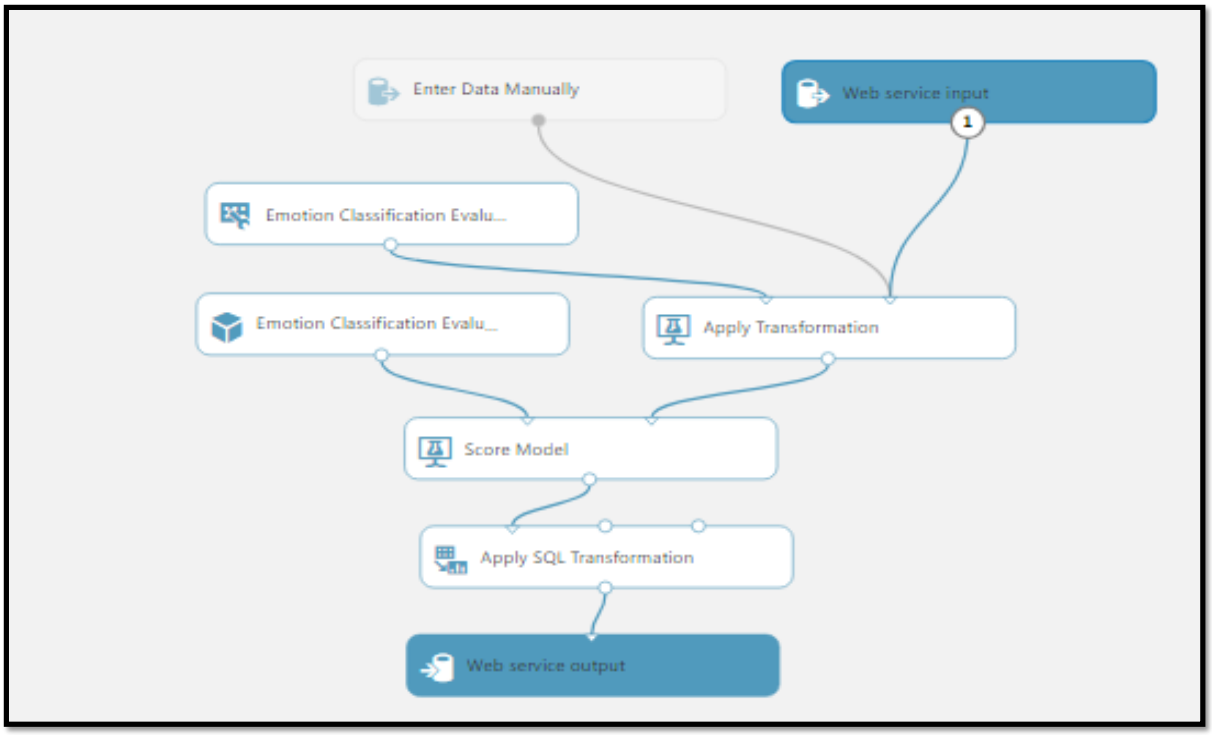

Figure 4.10: The back-end of the web based emotion recognition module setup

4.5 Results

A total of 1178 audio clips were collected for analysis. From the collected audio clips 70% were used for training, 15% were used for testing and the remaining rest of 15% was used for validation.

In the first phase of the experiment a neural network based methodology was used for classifying the affective state directly considering them to be discrete states rather than defining them through arousal-valance dimensions. This was implemented using a feed forward neural network with a single hidden layer with 25 hidden neurons trained to classify the affective states. Fig. 4.11 illustrates the confusion matrix for the classifier which uses spectral features. Fig.4.12 shows the confusion matrix for prosodic feature based emotion recognition system.Table 4.1 shows the performance of the two neural networks. From the performance of

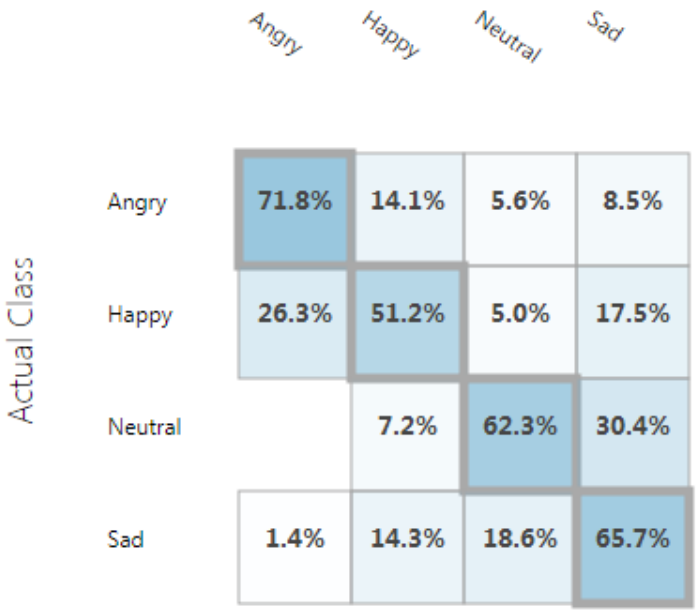

Figure 4.11: Confusion Matrix for Spectral Feature based Emotion Recognition Neural Network

Table 4.1: Performance Evaluation of Prosodic and Feature Based Neural Networks

Feature Type	**Overall Accuracy**	**Average Accuracy**
Spectral Features	62.41%	81.22%
Prosodic Features	83.33%	91.66%

these two networks prosodic features show better metrics from the two.

Hence *prosodic feature based neural network* will be used from here onwards to determine the emotion of the utterance.

4.6 Summary

This chapter presents an emotion recognition system which uses the vocal features of the user's response in order to determine the emotional state of the user. Vocal features are easier to work with than the visual features which makes it easier to do emotion recognition in real time. For this research the prosodic

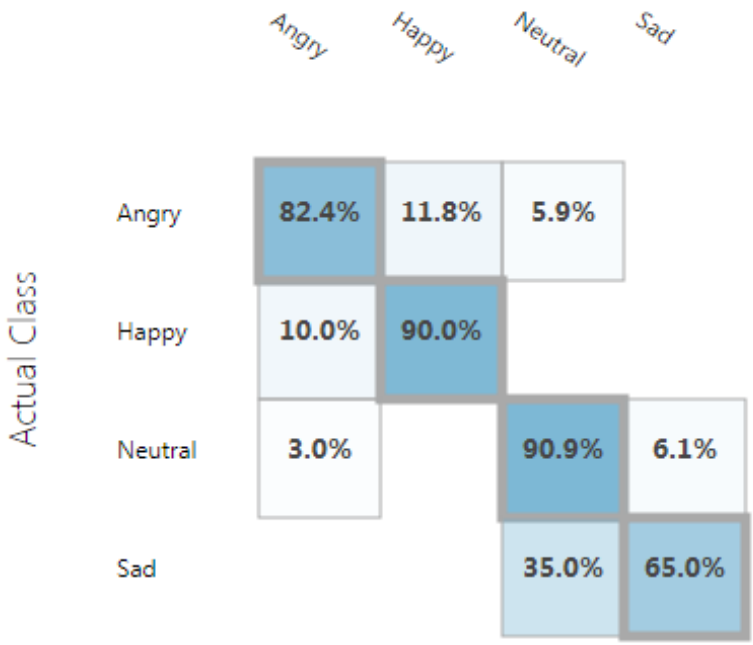

Figure 4.12: Confusion Matrix for Prosodic feature based Emotion Recognition Neural Network

and spectral features of the voice were used which are prominent in conveying emotional content of speech. Speech based emotion recognition is a relatively simpler, less time consuming and can be processed in real time compared to most of the other modalities. The presented system uses a neural network that is trained specifically to classify each of the considered emotions. This system was able to identify all the emotions with accuracies over 85%.

Chapter 5

LEXICAL ANALYSIS SYSTEM

5.1 Identifying Linguistic Message

The linguistic messages referred in this chapter corresponds to the text message from transcriptions of oral communication. The link between text and emotions has been studied from social psychologists to natural language processing researches. However during this research it was found that simply doing a sentiment analysis based on the text is not enough. Figure 5.1 illustrates the pitch and intensity contours for the utterance *"Will you tell me why"* in four different emotional states. Although the text based sentiment value is the same for all four instances in reality it does not properly convey the emotional state of the person. Therefore the proposed system uses both para-linguistic and linguistic information to analyze the emotional content.

Lexical analysis is important in order to identify the stimulus effect.Although the audio based emotion recognition system is used to find the emotional state, further analysis should be done to determine what triggered the emotion. Without knowing the triggering event the behavior modification cannot be done. Since this is a conversation system it is assumed that the action which stimulated in the emotion is presented in the previous turn.

The utterances that are presented by the user can be divided into following categories.

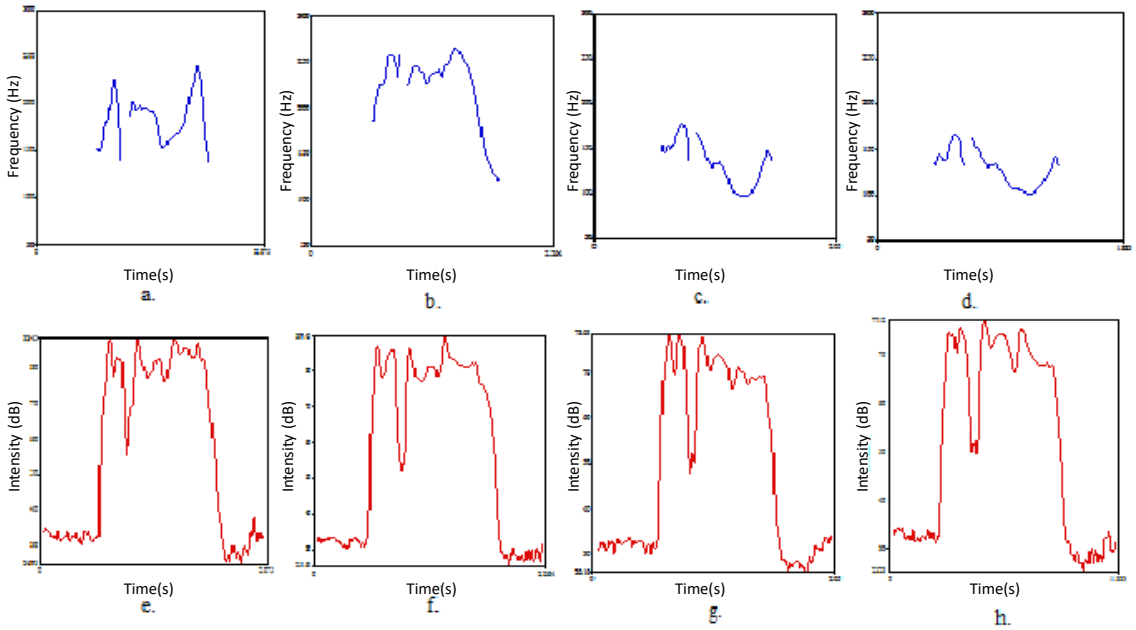

Figure 5.1: Pitch and Intensity Contours of a) & e)happy b)&f)angry c) & g)sad d) & h)calm emotions for the same utterance

- Statement - "She reads two newspapers every day " (subject + verb and optionally an object)
- Questions "What did you have for breakfast?"
- Command - "Open the door" (No subject)
- Exclamations - "What a beautiful painting!" relates with surprise emotion)

This system has the ability to process all the categories of the utterances except for commands. Commands are considered out of scope for the system of conversation since commands should be associated with physical behavioral tasks. However conversational commands such as "*Shut Up*" are considered separately as they have an immediate effect on the conversation flow.

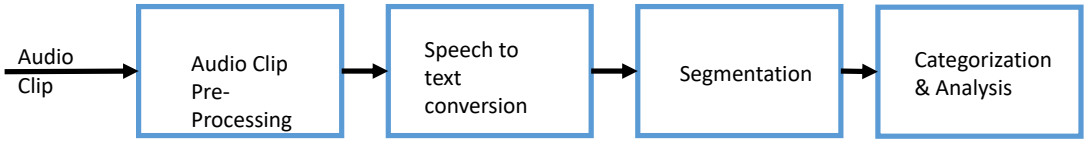

Figure 5.2: Flow Diagram of the lexical analysis system

5.2 Functional Flow of Lexical Analysis

Figure 5.2 illustrates the functional flow of the lexical analysis system. Each of the steps are described in detail below along with the relevant software tools.

- **Preprocessing :** The audio clip is first passed through the pre-processing unit where a series of filters are used to eliminate noise and other unwanted disturbances. A more rigorous pre-processing is done in this stage compared to the para-linguistic feature analyzer.

- **Speech to Text Conversation :** The next stage is a critical step where the speech is converted into text. The system uses both an on-line method with a much higher accuracy and an off-line method to be used in case an Internet connection is not available. In the practical implementation both Google speech and CMU Sphinx are used for this purpose.

- **Segmentation :** In the next stage of the system the utterance is tokenized and segmented into parts of speech. This is done by using the Google Text Analysis Service. Part of speech tagging is the process of marking up the word of the text transcription corresponding to the particular part of speech based on both definition and context. Part of speech tagging will be able to tag the most commonly used parts in English which are noun, verb, article, adjective, preposition, pronoun, adverb, conjunction and interjection. Fig.5.2 illustrates a part of speech tagging of an example utterance. Tokenization provide each word with a unique identification

My	school	principal	praised	me	for	my	decorations	at	the	art	festival
PRON	NOUN	NOUN	VERB	PRON	ADP	PRON	NOUN	ADP	DET	NOUN	NOUN

Figure 5.3: Parts of speech breakdown of the utterance

Figure 5.4: Part of Speech Segmentation

which is useful for further analyzing of the utterances as described in section 8.1.

- **Categorization & Analysis** In the final step the words are categorized into main entities and others.

  "My school principal praised me for my decorations at the art festival."

  Saliency is a measurement of how important the respective entity to the entire utterance. In this case the saliency is measured from 0 to 1 with 0 being least important to 1 being most important. The entity which has the maximum saliency is considered as the main entity which the memory is centered around. Apart from the saliency, sentiment relevant to the respective element is also considered. Table 5.1 present the entity level analysis of the example.

  In order to calculate the saliency a database of multiple utterances with saliency values for each entity are required for training. Such a database is

Table 5.1: Entity Analysis

Entity	**Category**	**Saliency**	**Sentiment**
School Principal	Person	0.43	0.4
Art Festival	Event	0.25	0
Decorations	Other	0.32	0

not easy to acquire manually, therefore Google lexical analysis is used. It has its own system trained using millions of snippets with saliency values.

The sentences used for the entity analysis are either simple or compound sentences.

- Simple sentence - Contains only on independent clause and no dependent clauses
- Compound sentence - Contains multiple independent clauses and no dependent clauses

Although the lexical analysis can be done for more complex sentences the emotion recognition system cannot handle an utterance with several clauses with sufficient accuracy. Hence the number of clauses do not exceed two. Further sentiment analysis can also be complicated with several clauses that convey different emotions.

5.3 Practical Implementation

The practical implementation of the lexical analysis was done through the MIVoIS Sentiment Analysis Module. Fig. **??** shows the Graphical User Interface of MIVoIS Sentiment Analysis Module.

- **Step 1** : The module allows the user to listen to the user's response and then stop when he stops. In the next step the user's speech response is converted into text. If the converted response is erroneous, the experimenter

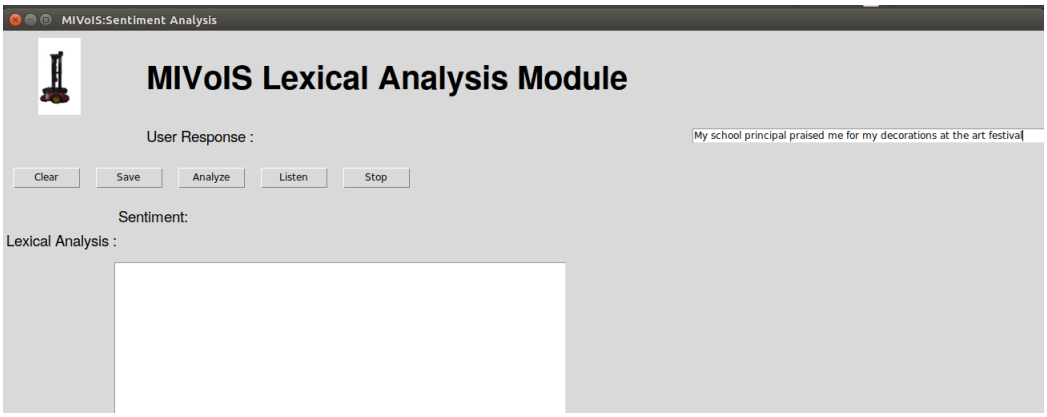

Figure 5.5: Step 1: Statement Input

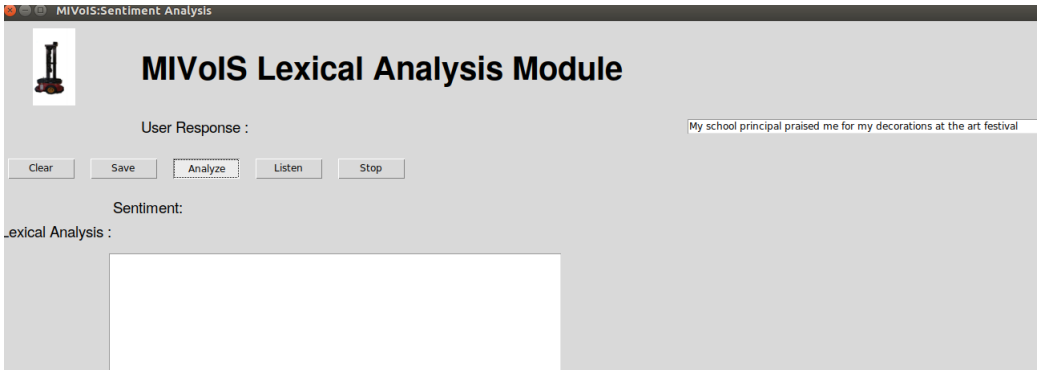

Figure 5.6: Step 2 : Analyzing the input

has the ability to change the response which appears as the output. This stage is shown in Fig.5.5.

- **Step 2** : In the next module analyzing of the user response is done. Within this module segmentation, categorization and analyzing is done in this phase. Fig.5.6 shows the GUI in this phase. This phase should be initiated by the experimenter by pressing the Analyze button.

- **Step 3** : As the final step the results are displayed which include the details of entities and their respective category, saliency and sentiment values. The final output is shown in Fig. 5.7.

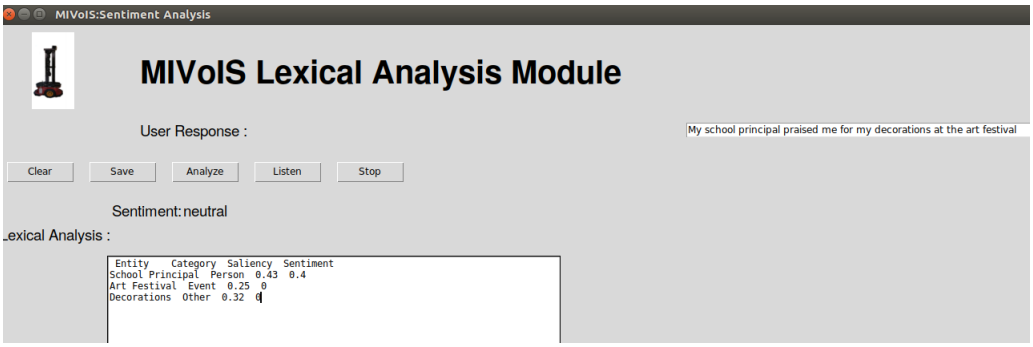

Figure 5.7: Step 3: System Output

5.4 Results

Data was collected by using both standard test databases and practically obtained audio clips using interactive sessions with the participants. As the test databases Surrey Audio Visual Expressed Emotion Database, Berlin Database of Emotional Speech and The Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS) [58] [59]. The practical data was collected by 15 participants who presented 3 utterances in the 4 affective states mentioned earlier.

All the utterances that were considered were simple sentences. Further the entire utterance corresponds to a single emotion and no utterances were used which include switching of emotions within the sentence.

To verify entity analysis 760 utterances were collected from the sources such as the participants utterances, test databases and other open source material which are freely available in the world wide web. None of these utterances comprised of more than 5 entities. Fig.5.8 shows how the entities are distributed among the utterances. The number of total entities in these utterances were 1866 of whom 1802 were correctly identified. In order to identify the most important entity of the utterance a panel of 5 individuals with a knowledge of English linguistics were chosen. They were given a printed version of the utterances along with the extracted entities from the system and were asked to choose the main entity expressed in the utterance. Afterwards these individuals were asked to listen the

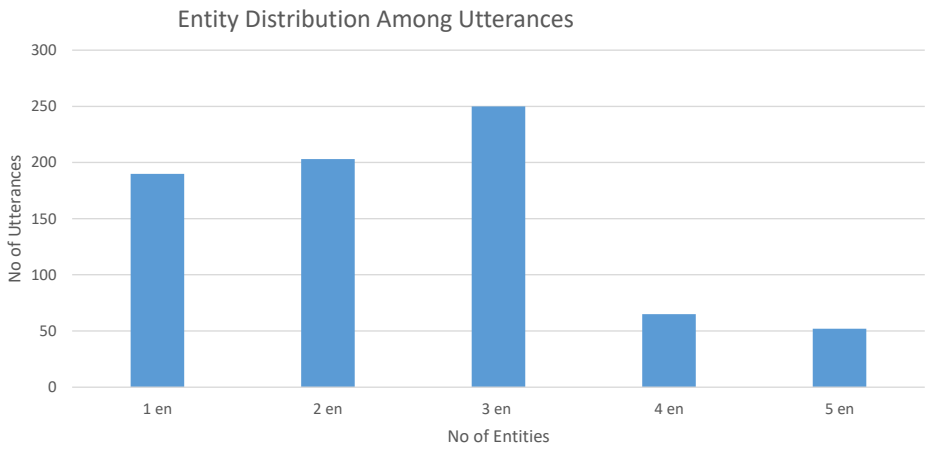

Figure 5.8: Distribution of Entities of Utterances

audio recordings of the utterances and identify the entities and choose the main entity. These two approaches were taken in order to determine whether speech techniques in order to highlight the main entity of an utterance can be adequately captured through linguistic methods. The main entity was chosen according to the preference of the majority of the panelists. In case of a tie any of the entities can be considered as main entities. The system was able to correctly identify 91.05% main entities that the panelists chose. When the panelists considered the audio recordings their responses changed due to the other vocal features such as vocal stressing of the words. From the 760 main entities 87.36% were chosen correctly when the vocal emphasis was considered.

5.5 Summary

This chapter presents the lexical analysis methodology that is used in the conversation system. In a conversation system it is important to understand what is being said and what are the important entities in the utterance in order to produce a response for the user. This conversation uses the concept of saliency and sentiment analysis in order to determine the main entity of the utterance.

By using this methodology the system will be able to determine the entity to which the emotions of the user are directed at. This is an important component of identifying the stimulus which triggered the emotion. The stimulus component is important when planning the response to the user as well as determining the emotional memories. This system is an addition to the speech feature based emotion recognition system in order to determine the stimulus of the emotion through a lexical analysis for the utterance.

Chapter 6

EMOTIONAL MEMORY DETERMINATION

6.1 Defining Emotional Memories

Emotional memories are defined as instances which have an emotional significance to the user. During conversations people convey their opinions about different entities they come across. For a robot to be able to have a meaningful conversation with a human it is vital for it to be able to differentiate the entities. Emotional memory module is used to retain the emotional memories. The emotional memory consists of two components which are namely entities and emotion. The emotion is detected through the vocal parameters of the speech . In order to extract the entities, the speech response of the user is first converted into text. Anything included in the utterance with a distinct and independent existence is considered as an entity.These entities are then extracted from the converted text which are divided into person, event, location, consumed goods or other entities. The other category comprises of the significant entities which do not belong to any of the above mentioned categories.

Apart from the entity related emotional memory there can be behavior related emotional memory. Behavior related emotional memory consists of responses of the users corresponding to the robot's behavior. Behavior based emotional memory may or may not include entities and are hence excluded from this study.

6.2 Human Emotional Memories Labeling Process

Long term memory required for a robot to be a long term companion can consist of the following components [60].

- Episodic Memory - Personal experiences
- Semantic Memory - General factual information
- Procedural Memory - Task performing procedures

From these three categories the emotional contents are closely related to the episodic memory regarding the personal experiences. The personal experiences can be robot's own or any of the user's the robot is associated with. Humans use a process known as emotional labeling in order to encode memories related to their lives [61] [62] [63]. By the labeling process humans will be able to recall the past experiences and reminisce the important points learnt from the experience and the emotions felt [61]. For the robot to successfully converse with the user it should have its own episodic memories along with the user's memories. This research proposes a method for the robot to develop its own emotional memory regarding the user. The current systems are mostly theoretical and is more focused on the robot's own emotional memory [61]. Yet if the robot is to develop conversational capabilities to be par with a human companion it should be aware of the emotional affiliations of the user with certain objects, events, locations or other entities. It is difficult for all these information to be preprogrammed and has to be learn t through the conversations as these emotions can change through the course of the time. This model can be used to create emotional memory relevant to both the robot and the user it is associated with.

6.3 Determining Emotional Significance

During conversations people convey their opinions about different entities they come across. For a robot to be able to have a meaningful conversation with a human it is vital for it to be able to differentiate the entities. Emotional memory module is used to retain the emotional memories. The emotional memory consists of two components which are namely entities and emotion. The emotion is detected through the vocal parameters of the speech . In order to extract the entities , the speech response of the user is first converted into text. Anything included in the utterance with a distinct and independent existence is considered as an entity.These entities are then extracted from the converted text which are divided into person, event, location, consumed goods or other entities. The other category comprises of the significant entities which do not belong to any of the above mentioned categories.

Apart from the entity related emotional memory there can be behavior related emotional memory. Behavior related emotional memory consists of responses of the users corresponding to the robot's behavior. Behavior based emotional memory may or may not include entities and are hence excluded from this study.

Figure 6.1 illustrates the component of the memory module. The memory module has two components with one considering the emotional memories relevant to the robot and the users who interact with the robot. In this system however the focus is on how the robot can handle emotional memories related to the user. Emotion and memories have a strong link in humans as they tend to recall events with clarity in a more detailed manner than any neutral or routine event. Hence if the robot is going to act as a companion it should be able to recall the certain events that are significant for the user. Unlike the humans robots do not have any hard criteria to differentiate between events which have an emotional significance and neutral events.

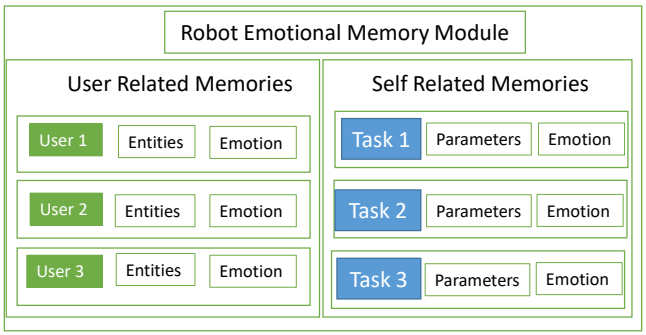

Figure 6.1: Representation of the memory module

In this research only the user related memory component is considered. The robot can be expected to interact with several members of the family in a domestic environment. Therefore for each member of the family the robot should maintain an emotional memory profile. In this memory module the robot will store the entities that are relevant to the user. There can be several entities which correspond to a single memory. In this system only memories associated with a single emotion are considered.

The overall emotional memory differentiating process is illustrated in Fig. 6.2. It is important to note that the entity level analysis is done only if the utterance is considered as emotionally significance. Otherwise the utterance is seemed as a routine event and hence considered not relevant for considered storing in the long term memory. The utterance is considered emotionally significant if the detected memory is one of happy ,angry and sad.For an example one of the users uttered the following sentence regarding a school experience:

> "My school principal praised me for my decorations at the art festival."

This is a recall of a positive experience of a participant. The emotion detection unit recognized this as a happy memory according to the classifier output. For analyzing the effect of the entities to the memory saliency and sentimental value of the entities are taken into consideration.

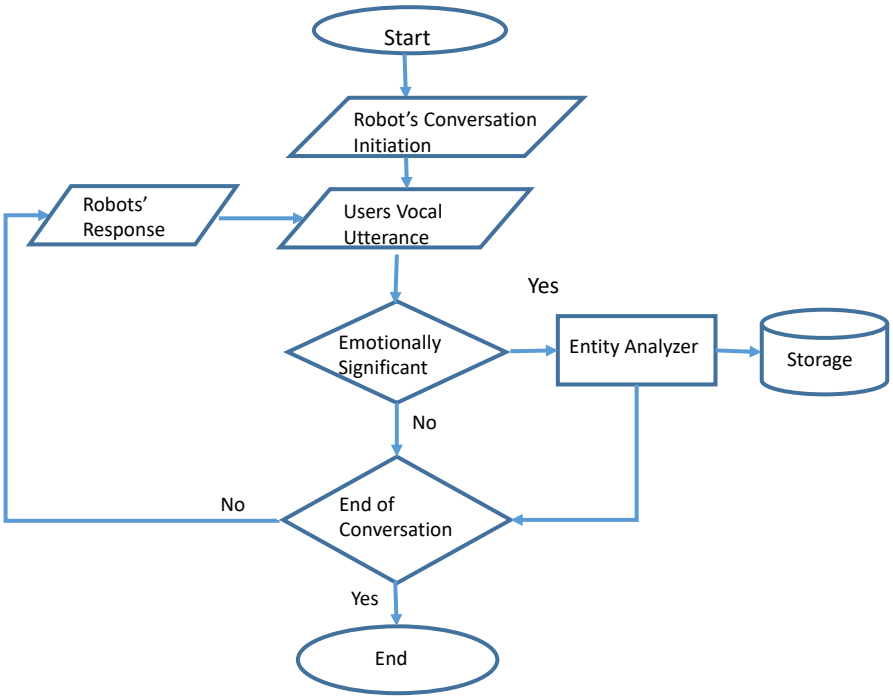

Figure 6.2: Overall System of Emotional Memory Detection

According to the entity level analysis of the utterance this incident can be considered as a happy memory related to the school principal involving decorations and art festival. The main entity that this memory is centered around is the school principal which belongs to the person category. Hence the system can concludes that the school principal is a person whom the user has a positive perception. Fig.5.3 illustrates part of speech breakdown of the utterance. The entities of this utterance consists of compound nouns school principal and art festival along with noun decorations.

6.4 Results

A total of 15 participants were used for the experiment. In this experiment the the participants presented four instances where two of them had an emotional significance and two other neutral events. The robot first analyses the audio parameters to check whether the utterance is emotionally intense or not. If the

classifier chooses one of the emotions then the robot will proceed on to determine the entities that are associated with the emotion. The users responses regarding the instances were compared with the robot's identification. Although compound nouns were used for this experiment, the nouns which can function as both a noun and a verb are not used.

The performance of the emotion recognition system Table 6.1 presents a portion of the results that were recorded during the experiment of 3 participants. After the experiment the robot updates the entities and their emotional affiliation of the user's profile. It is also important to note that the same main entity can give out different emotional reaction when the associated entities differ.

The emotion recognition system was able to detect the emotion in 48 times out of 60 instances. From all the instances the system was able to correctly detect 22 instances out of the 30 emotionally significant instances. The neutral state recognition performed better identifying 26 of the 30 instances.

In the entity identification module the most critical segment was speech to text conversion which directly affects the entity identification process. The speech to text conversion happens only if the utterance which is selected is considered as emotionally significant. The speech to text conversion correctly captured the entities of 19 out of the 22 identified emotionally significant instances. Out of all the properly captured utterances the system captured all the entities expressed in these utterances.

This system enables the robot to build a profile on its own of a certain user and regarding the nature of relationship of entities. These entities can be used as keywords in response generation with regards to emotions. Therefore the human-robot interaction can be enriched by the robot being able to relate to the user on a more personal level. This system can also be used in-order to determine the user perception of the entities which can be useful when the user is conversing with the robot.

Table 6.1: Comparison of the Participants Response vs The Robot's Response

Person ID	**Instance No**	**Person's Response**			**Robot's Identification**		
		Emotion	**Main Entity**	**Related Entity**	**Emotion**	**Main Entity**	**Related Entities**
1	1	Sad	Puppy	Sicknesss	Neutral	-	-
	2	Neutral	Market	-	Neutral	-	-
	3	Neutral	Shoe	-	Neutral	-	-
	4	Happy	Exam	Results	Happy	Exam	Results
2	1	Angry	Man	Garbage, Front, House	Angry	Garbage	Front,House
	2	Neutral	Cricket Match	-	Happy	Cricket Match	
	3	Happy	Car	-	Happy	Car	-
	4	Neutral	Dinner	Bread	Neutral	-	-
3	1	Neutral	Office	Meeting	Neutral	-	-
	2	Neutral	Kitchen	Biscuits, Jar	Neutral	-	-
	3	Happy	Dog		Happy	Dog	
	4	Happy	Uncle	Australia, presents	Happy	Uncle	Australia, presents

Further it is recommended that in the future the robot to be able to determine the entities in complex sentence where multiple emotions are involved. In some cases it was observed that that multiple emotions are being conveyed through speech. This is due to the fact that there can be happy, sad or angry instances of the routine activities which can bear no importance in long term memory.

Another limitation of this system is not being properly able to identify the entities that can either be considered as a noun or a verb. The ambiguous nature of such instances can cause the system to not detect the relevant entities of the utterance. Overall this system can be considered as a starting point for the robots to extract the entities with emotional significance of the user which will fulfill their quest to be longterm companions of the users. Since human perception of the entities change over the time and new entities are introduced through course of the time it is important for the system to update accordingly.

6.5 Summary

In a conversation system there are many variables that should be retained in order to converse with the user. It is important to retain this information as it contributes to finding the stimulus event and for further analyzing which is required to make conversation decisions. This module uses natural language processing libraries together with the linguistic tools in order to extract the entities and to identify main entity in the utterance. The proposed system will be able to identify the most important elements of the utterance, which will make the storing of the user's memories more efficient. The main limitation of the system is not considering vocal elements such as emphasis or stress in identifying the main entity. Finally it is recommended that the intensity level of the emotion should also be considered when deciding on emotional significance.

Chapter 7

CONVERSATION DECISION MAKING

7.1 Conversation Rule System

The dialog system that is been used for this experiment is made specifically for two party conversation. Therefore the simple turn taking rule applies, in which

- The human and the robot takes turns one after another.
- The amount of silent time between the turns is kept as minimum as possible. In case the dialog system takes excessive amount of time to produce a response the human operator interferes with the operation and propose a response.(This is done in order prevent the user from misinterpreting the silence as a significant silence which can be interpreted as refusal to respond.)
- The dialog system has a set of predefined questions belonging to a set of topics chosen prior to the experiment after discussing with the participants.
- The questions are presented in an order until the system detects that the participant has lost the interest in the conversation.
- Then the system can either change the topic or abort the conversation depending the level of enthusiasm.

7.2 Decision Making on Continuation of the Conversation

One of the most important decisions of a conversation is to determine when the conversation is going to end or else whether it can be continued or whether the topic of conversation should be changed. For humans this change is done using either his/her preferences or by observing the body language of the participating parties. Humans have the ability to determine the willingness to interact by observing and listening to various body language and vocal features.

7.2.1 Interaction Decision Determination

The affective state for the respective turn of the conversation is stored in a database. It is used to determine whether the affective states are persistent or temporary. If the affective state is seen in three of the last four turns the affective state is considered persistent or else it is considered as a temporary state. The rules that are used in the conversation decision unit are given in Table 7.1. These rules are applied until all the system has covered all the three topics and come to the end of the conversation.

Happy and neutral states are considered as positive states which shows that the user is interested the conversation. Angry and sad states are considered negative states. Hence those states indicate lack of interest in the conversation. It is important to note that the system immediately changes the topic when angry state is detected and hence a persistent angry state does not exist. Although sad state is considered negative, behavior modification does not need to happen unless the state is persistent.

Table 7.1: Rules System for the conversation Decision System

Affective State	**Persistent or Temporary**	**Action**
Happy	Persistent	Continue the conversation
Happy	Temporary	Continue the conversation
Neutral	Persistent	Continue the conversation
Neutral	Temporary	Continue the conversation
Sad	Persistent	Topic Change/Termination
Sad	Temporary	Continue the conversation
Angry	Temporary	Topic Change/Termination

7.3 System Overview

The functional flow of the proposed system is given in Fig. 7.1. This methodology is used to evaluate the user engagement level of a human-robot vocal interaction system. Both temporary and persistent emotional states are analyzed by the system. The final objective of the system is to change the topic of the conversation if the user engagement is at a low value persistently. The system takes the audio waveform file of the users utterance as the input for the system. The system then feeds this audio waveform separately into the vocal feature and the lexical feature extractor. The extracted vocal and lexical features are the passed to the emotional classifier.

The emotional classifier will use a neural network based system to determine the emotional state. The lexical feature analyzer will also determine the key phrases of the users utterance. The emotional state and the key phrases are then stored in the data recording system. The data is then passed to the analyzer which will compare the received data to historical data and decide whether the emotional state is temporary or persistent.

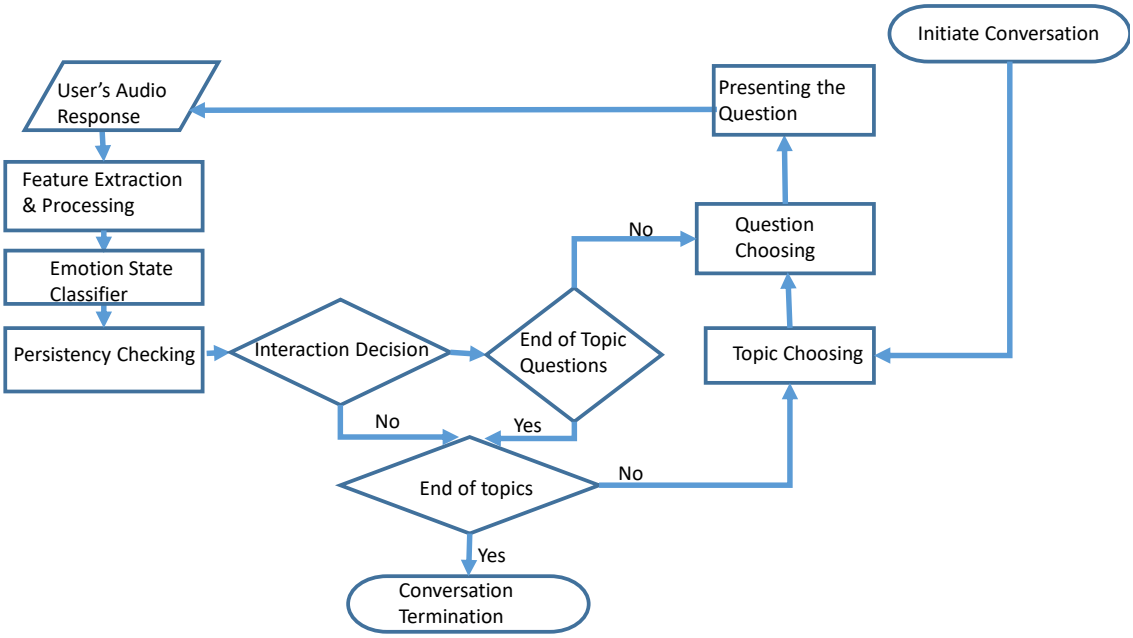

Figure 7.1: Functional flow of the system implementation

7.4 Implementation of the System

The concept has been implemented on MIRob platform with a Microsoft Kinect sensor attached [64]. The required navigation maps are created with Mapper3 Basic software. The navigation maps are used to keep the set social distance required for the conversation which is always 1 m away from the user. The audio clips were captured using the array of microphones in the Kinect. Further a voice synthesis unit is used to produce the vocal response from robot. The experiment was carried out in a simulated domestic environment. Fig. 7.4 illustrates a participant engaged in vocal interaction with the robot during the experiment. The robot and the participant were kept at a distance of 1m apart facing each other. Prior to the experiment each of the participants were given the opportunity to pick 2 topics from a given pool of topics. Each of these topics have a set of 10 questions associated with it. Fig. 7.3 illustrates a question set that was used in the experiment. Once the topics are chosen they are given as inputs along with the vocal interaction system of the robot. The system will choose another random

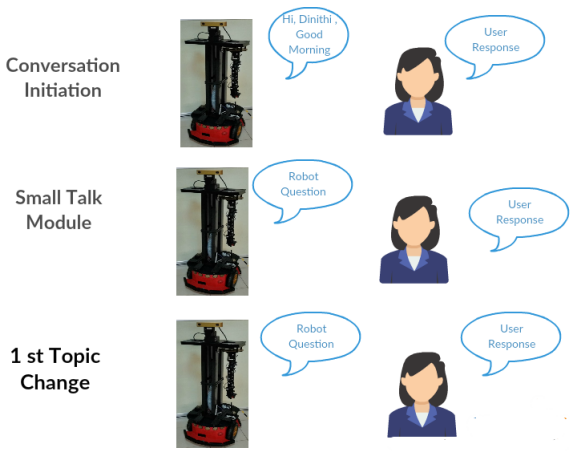

Figure 7.2: Conversation Initiation of the Robot and the User

topic from the rest.

The system will initiate the conversation with a friendly greeting. The figure 7.2 shows how the robot will initiate the conversation with the user when the user is new. The small talk module is designed so that the robot and user will introduce each other to make each other familiar with the other party. In the small talk module it is important for the robot to be consistent with its own personal information. Therefore the robot uses a rule based conversation system for ensuring consistency. In case of the user interacting with the robot not for the first time the robot will skip the small talk module and go directly towards the 1st topic change. The robot will start randomly with one of the 3 selected topics. The robot and the participant will take turns one after the other until the conversation decision system indicates the user has lost interest of the conversation. When the user has lost the interest the system will switch to a new topic. This process happens until the all the 3 topics are used in the conversation. Finishing all the questions result in automatic termination of the conversation.

Topic :Animals

1. Did you grow up with pets in your home?
2. What do you think is the best pet to own?
3. Do you have a pet?
4. What can children learn by having a pet?
5. Are there any animals that you are afraid of?
6. Which animal do you think is the most intelligent?
7. If you had to choose one animal to be, which one would you be?
8. What animal ability would you like to have like flying or breathing underwater?
9. What is your favorite memory with a pet?
10. Have you ever been bitten by an animal?

Figure 7.3: Question set used for the topic animals

Figure 7.4: A participant and MIRob platform during the experiment

Table 7.2: Results of the Experiment

User No.	**No of turns before change/termination**			**User Reaction**			**Overall System Satisfaction**
	Topic1	**Topic 2**	**Termination**	**Topic1**	**Topic 2**	**Termination**	
1	6	4	8	satisfied	unsatisfied	satisfied	satisfied
2	5	5	7	satisfied	satisfied	unsatisfied	satisfied
3	4	6	6	satisfied	satisfied	satisfied	satisfied
4	6	8	5	satisfied	unsatisfied	satisfied	unsatisfied
5	9	5	7	satisfied	satisfied	satisfied	satisfied
6	7	5	6	unsatisfied	unsatisfied	unsatisfied	unsatisfied

7.5 Results

A conversation system should be able to make decisions regarding the flow of the conversation. The system proposed will make decisions on continuing the conversation, changing the conversation topic and termination of the conversation. In this systems these decisions are taken with considering the emotional state of the user.

The experiment was conducted with the participation of 20 individuals in the age range of 18-58 (Mean-30.8 and Standard Deviation-11.47). Before the experiment 12 audio samples were obtained from each of the participant with 3 each displaying the 4 affective states mentioned in the previous section. Throughout the conversation the user had to converse about 3 topics. Within the conversation there will be two topic changes and finally the termination of the conversation. After the conversation each of the participant was asked whether they are satisfied about when the changes and termination happened in the conversation.

Figure 7.2 shows a portion of the results which include 6 users.A combined total of 60 topic changes and terminations happened through the experiment. The average number of turns a participant had was 6 per topic. Further the average number of participant turns for the topic chosen by the system was 4.65 turns per topic. Of all the changes and terminations 57.01% of them was considered satisfactory by the participants. A total of 16 participants of the total considered the overall system to be satisfactory. The figures 7.6 and 7.5 shows the user satisfaction when the system uses the decision making with and without considering emotions.

One of the reasons which might have caused unsatisfactory performance is due to the fact anger and happy states having similar acoustic properties. Although these two states can be easily distinguished through visual means such as facial emotion recognition, in acoustic measures the high excitement and high annoy-

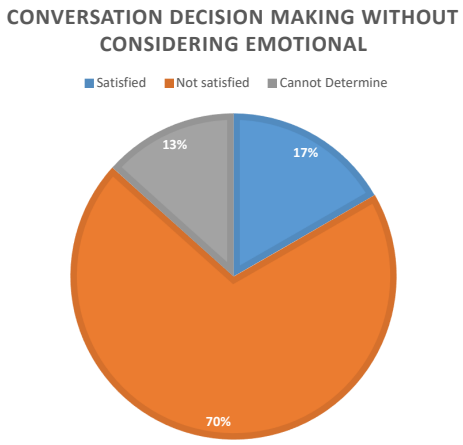

Figure 7.5: User Feedback on Conversation Decision Making without using emotional states

ance states both have nearly similar values for prosodic parameters. Hence it is highly likely that these two states being misclassified which can produce two completely different results in the system.

Further the use of a static value of 3 out of the last 4 turns being the same affective state considered as a persistent affective state should change according to the individual. The participants suggested that this value does not provide a satisfactory performance when the conversation is about an unfamiliar topic such as the random topic generated by the system. The user had to carry on the conversation for at least 4 turns with a negative affective state in order to change the topic which participants suggested as a lengthy conversation than expected on a topic not favorable for conversation.

7.6 Summary

The subsystem described in this chapter uses the emotional state of the user in order to make decisions regarding the flow of the conversation. The decisions that

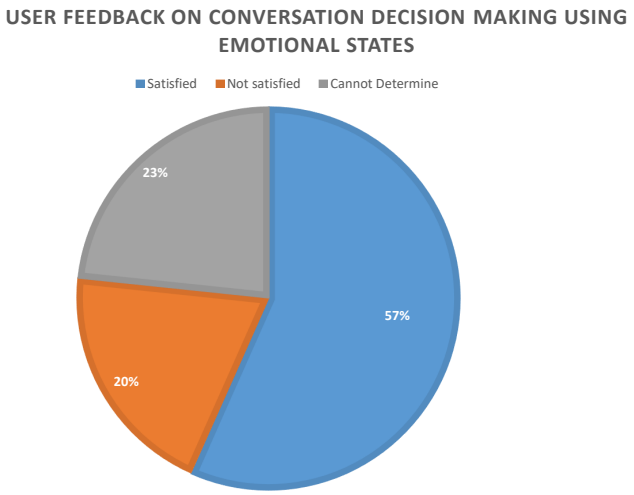

Figure 7.6: User Feedback on Conversation Decision Making using emotional states

are made during the conversation flow are conscious of the emotional state of the user. Further the robot is also aware of terms associated with the termination of the conversation which takes priority over the emotional states. This method will provide a basis for conversation flow decision making, is generally lacking from the traditional artificial dialog systems. Further it can be clearly seen that the user's satisfaction increases when the system takes the emotional state of the user when decisions regarding the conversation flow are made. This can be considered as an improvement for the traditional conversation system.

Chapter 8

VOCAL RESPONSE GENERATION

8.1 Definition of Empathetic Responses

The conversation system used by the robot uses its sensors to determine the emotions of the user and adapt the conversation according to the user's emotion. The system is designed to display empathetic behavior towards the user. In order to show empathy the robot's behavior function B_r should be a function of the user's emotional state (es) as shown in .8.1. Further in order to give a personalized experience the robot should consider the personal profile of the user (up).

$$B_r = f(es, up) \tag{8.1}$$

The emotional states considered for es are happy, angry, fear, disgust, sad and surprise. The conversation system has a set of behaviors that can be specified for each emotion. The conversation between the user and the robot is made up by consecutive turns.

The conversation system uses a set of social norms that are prevalent in the society to assess the stimulating event of the emotional state of the user.

Table 8.1: Possible Stimulus Events and Expected Behaviors

Emotion	**Stimulus Event**	**Expected Behavior**
Happy	Gaining of a valuable object Reminising a delightful memory	Sharing the happiness
Angry	An unplesant event An objection	Managing Anger
Fear	A threat	
Disgust	An unpleasant object, event or a person	
Sad	Loss of a valuable person or an object	Consoling the person
Surprise	An unexpected event	

8.2 Factors Affecting Empathetic Behavior

In order to simulate empathetic behavior there are elements which should be considered other than the user's emotional state. There are several factors to be considered when presenting responding behavior towards the user which are described below.

8.2.1 User Preferences

One of the important aspects of showing empathetic behavior is to show the user that robot understand how the user is feeling which helps in creating a bond between the two required for long term interaction. Therefore it is important for the robot to have a basic understanding of the user's preferences. There is no need to have a complete overview on the user at the start. The user profile can be updated along with the time as personal preferences change over time so an adaptive approach is required from the robot. Apart from the basic personal information that the robot requires to know. Much more intimate knowledge about the user such as favorite song which cheers up the user can be used in showing empathetic behavior.

In case of robot not having enough data on a user profile, the system can compare with the other users with similar available data and make reasonable

assumptions for the missing data. Further when the robot run out of empathetic behavior for one person the user profile reference of other similar interest people might be useful to the robot to get suggestions of novel behavior.

8.2.2 Effect of the Previous States

The previous behavior that the robot displayed is important for the robot to properly plan the next behavior. This will ensure that the robot will not be performing any repetitive behavior one after another. This condition is very important when the emotional states are permanent. The robot maintains a stack of all the results of called behavior functions.When calling new functions precautionary methods should be taken to ensure that there is sufficient distance between two alike states chosen by the algorithm.

The knowledge about previous behavior is also important to keep the user engaged in the conversation by introducing novel behavior. Especially the users will not appreciate if the robot is displaying the same behavior over and over again for the same emotional state. The user's expect novelty as well as consistency of the information at the same time.

8.3 Stimulus Events and Expected Behavior

After the detection of the emotional states the next task of the robot is to show emotional response through conversation. In the practical implementation the only output modality available is voice. Therefore separate voice profiles were created which corresponds to the emotional states that the robot is able to detect. The voice profiles were created by changing parameters such as speech rate and volume.The spoken response also was chosen from a corpus that was specifically developed for the system which included transcripts of the previous interactions of the robot as well as by studying real world scenarios where similar

human-human interactions are observed .

In case of a conversation the stimulus event is always assumed to be a what was said. Therefore in order to analyze the stimulus event of the this system uses the lexical analysis which was described earlier.

8.4 Vocal Response Generation System

After the detection of the emotional states the next task of the robot is to show emotion. In the practical implementation the only output modality available is voice. Therefore separate voice profiles were created which corresponds to the emotional states that the robot is able to detect. The voice profiles were created by changing parameters such as speech rate and volume. The spoken response also was chosen from a corpus that was specifically developed for the system which included transcripts of the previous interactions of the robot as well as by studying real world scenarios where similar human-human interactions are observed .

In order to generate empathetic behavior the architecture used in this research is corpus based. In this method the system uses a recurrent neural network in order which maps the user's turn to other user response. The following steps are followed in the response generation system.

- Acquiring the data sets
- Creating the model
- Training the model
- Testing the model

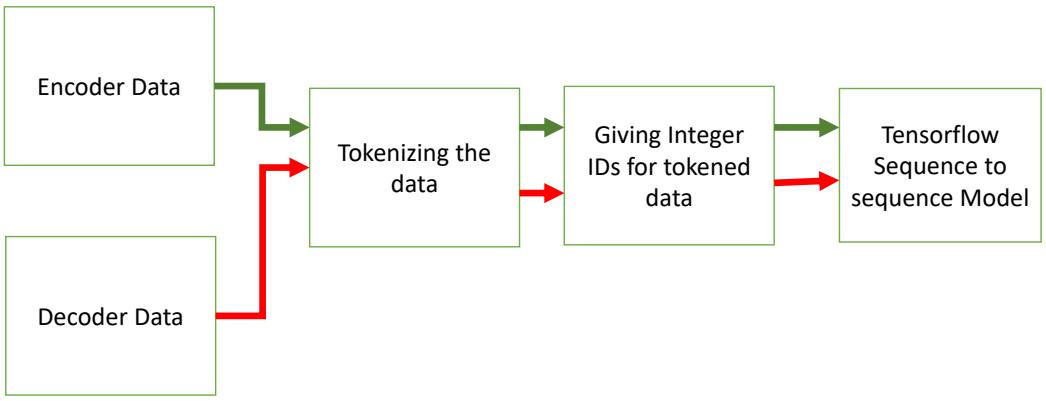

Figure 8.1: Feeding encoder and decoder data into the Model

8.4.1 Implementation

Acquiring and Processing Data

In order to implement the system there was no special need of hardware and the existing robot platform and the audio system was used. For the first part in acquiring data sources such as the databases from Cornell University Movie Corpus, Kaggle, Ubuntu Dialogue Corpus as well as conversations from the participants were used. These acquired data was divided into two categories.

- Encoder Data - Utterance from one side of the conversation
- Decoder Data- Response of the other side

After categorizing the data the encoder and decoder data are tokenized separately. The tokenized data is then given Integer IDs for unique identification. This tokenized encoder and decoder sets are then fed into the model. This process is illustrated in Fig. 8.1.

Creating and training the model

The model used for vocal response generation is the sequence to sequence model. This model is frequently used for deep learning conversational systems. This model consists of two recurrent neural networks. Encoder maps a variable-length source sequence (input) to a fixed-length vector. Decoder maps the vector representation back to a variable-length target sequence (output). Two RNNs are trained jointly to maximize the conditional probability of the target sequence given a source sequence.

Each hidden state influences the next hidden state and the final hidden state can be seen as the summary of the sequence. This state is called the context or thought vector, as it represents the intention of the sequence. From the context, the decoder generates another sequence, one symbol(word) at a time. Here, at each time step, the decoder is influenced by the context and the previously generated symbols.

In this subsystem the sequence-to-sequence model is implemented through the Tensorflow library. The tokenized encoder and decoder data are fed into the modal for training. The system produces the response with one word at a time. The next word is always dependent on the previous word produced hence a recurrent neural network based system is the most suitable.

8.5 Results

In the first part of the experiment the system was tested separately for the response generation by using a testing sample of 250 utterances. These utterances and generated responses were given to a panel of 5 participants in order to assess the suitability of the response to the utterance. The participants were to assess the generated responses based on their relevance and meaningfulness. Further the

Table 8.2: Results of User Assesment of the system generated responses when emotions are not considered

Not Considering Emotions	**Meaningful**	**Not-Meaningful**
Relevant	16	-
Not Relevant	67	177
Total	83	177

participants were asked to present their own responses for the given utterances. These model responses were used to train the model. The model has the ability to train using the new data which enables it to modify the responses over the time to improve the performance of the system. The system was tested in two stages as given below:

1. Not considering the emotional states of the user

2. Considering the emotional state of the user

The purpose of using a two stage system is to assess the the effect of the emotions towards the response generation. Relevance in this experiment refers to how suitable the response is to the given situation. Meaningfulness means whether the generated words can convey a meaning to the other person. In order to be a successful response it should be both high in relevance and meaningful. Therefore for the assessment of the relevance, the response is first considered for meaningfulness.

The overall conversation system was tested with the participation of 15 users. Each user conversed with the Mirob for at least 5 turns. This is the final output of the dialog system. One conversation of the experiment is given below between Mirob and the user. It is important to note that the assistance of a wizard was used in cases where the users' utterance was not correctly captured by speech to text converter for better performance.

Given below is a conversation between a user and MiRob. In this conversation in case the speech to text converter was not able to correctly identify the utterance

Table 8.3: Results of User Assesment of the system generated responses when emotions are considered

Considering Emotions	**Meaningful**	**Not-Meaningful**
Relevant	34	-
Not Relevant	27	189
Total	61	189

an operator was used to identify the correct utterance.

MIROB : Hi, I am MiRob

USER1 : Hi, I am Chapa (MIROB *detects happy emotion and proceeds with the conversation*)

MIROB : Nice to meet you. What are you doing today?

USER1 : I came to the lab for experiments. (MIROB *detects neutral emotion and proceeds with the conversation*)

MIROB : Good I have lab friends.

USER1 : Nice to know that.(MIROB *detects neutral emotion and proceeds with the conversation*)

MIROB : You nice too

USER1 : Ok then (MIROB *detects neutral emotion third consecutive time and change the topic*)

MIROB : Do you like dogs?

USER1 : Yes, I have several dogs at home.(MIROB *detects neutral emotion and proceeds with the conversation*)

MIROB : Dogs come home meat like

USER1 : My dogs eat meat and fish.(MIROB *detects neutral emotion and proceeds with the conversation*)

MIROB : Dogs good animals now.

USER1 : Ok I have to go Mirob (MIROB *detects intention of termination and*

Table 8.4: User Utterances and Generated Response

Users Emotion(Detected)	**Users Response**	**Robots Response**	**User Comment**	
			Relevant	**Meaningful**
Happy	We are going to the South for a vacation	Great have good vacation	Yes	Yes
Sad	My puppy is very sick.	Sorry my puppy	Yes	Yes
Neutral	I went all the way to the market	Market very busy	Yes	Yes
Angry	He cannot carry on with that work	I am sorry him carry	No	Yes
Happy	Finally they stopped fighting over on what movie to watch	I not fight movie	No	No

Table 8.5: The comparison of system generated and human specified utterances

Users Utterance	**Humans Response**
We are going to the South for a vacation	Wow, have a great vacation
My puppy is very sick.	Oh, I am very sorry. What happened to the puppy
I went all the way to the market	Well sometimes you can't buy anything without going to the market
He cannot carry on with that work	Of course, you have to remind him the rules
Finally they stopped fighting over on what movie to watch	What a stupid fight.

proceed to terminate the conversation)

USER1 : Bye. Have a nice day.

Although the responses of the conversation system most often do not produce a grammatically correct response, the users will be able to figure out the meaning of the utterance by using the keywords of the system's utterance. Hence with further training of the system meaningful responses can be expected from the system.

8.6 Summary

The final goal of this system is to produce a vocal response to the user's utterance. In order to generate a vocal response a combination of recurrent neural networks are used. This approach increases the possibility of producing a response that is both relevant and meaningful compared to conventional systems. Further this systems provide more flexibility over the traditional rule or pattern matching systems. This system increases the overall satisfaction of the user and has the potential to improve its performance over the time with the new data. Further the system is also able to keep the consistency over the time regarding personal information of the user by including a pattern-matching and rule based

system. This hybrid system can be seen as an improvement over the traditional systems as it is able to keep strengths of both rule based and the learning system. Further this system has the ability to improve the performance over the time by using the transcripts of new conversations.

Chapter 9

CONCLUSIONS

This research proposed a conversation system for a domestic service robot which can use the emotional content of the user's voice in order to generate the responses and to make decisions regarding the continuation or termination of the conversation. This system can be described as a step towards emotional intelligence for service robots. Though emotion recognition systems are currently common in robots and artificial agents, it is rarely seen that a system uses such information in order to make decisions. Humans use emotions for decision making in social interactions. Displaying and understanding emotions is an important factor if the robots should be considered as companions in the future.

An emotion recognition system based on paralinguistic features of speech was developed for this system using an ensemble of neural networks. In order to understand the stimulus event of the emotion, lexical analysis was introduced. The lexical analysis system is also used to determine whether the user utters terms which indicate the termination of conversation. By combining emotion recognition and lexical analysis, a method to determine emotional memories and determine emotional stimulus was proposed.

9.1 Overall Assessment of the System

The system described in this research can be used for emotion recognition and to identify the stimulus event which caused the emotion. The emotion recognition

system uses the vocal speech features to extract the emotional content of the utterance. This system is able to determine seven emotions, although all the emotions are not considered in the decision making process.

A method was introduced to estimate the attention level of a person who is engaged in a conversation with a human-robot vocal interaction system. The main improvement of this system over the existing systems is that it takes the decisions based on the users affective state. For further developments, using a multi modal approach by including a visual channel, the system will be able to determine the attention level of the user when both listening and speaking.

A response generation system based on emotional state was developed using recurrent neural networks. This system is an improvement from the current hard coded systems which do not consider the emotional state of the user. The system which can use emotional states can produce more relevant responses The system needs to be trained with a much larger set of conversations. With more training data sets the system will be able to generate more relevant and meaningful responses.

To support the above systems an emotional memory system was also introduced in order to store the emotional experiences that the system gets to hear from the user. These data can be used for further analyzing to understand the preferences of the users in a detailed manner. The further analyzing of the emotional memories is not discussed in this research.

9.2 Limitations of the System

The system consisted of several subsystems each dedicated for a specific function. Each of these systems are mostly reliant on the output of the previous subsystem. However there are several limitations on the system which hinders the performance of the overall system.

This system when using individually trained networks for emotion recognition is able to produce results of higher accuracy compared to the single classifier emotion recognition. This higher accuracy comes at the cost of much longer processing speeds. This might not be desirable for dialog systems as longer the time it takes for the robot to produce the response the other party can get uninterested in conversation or may misunderstand as the end of the conversation.

Further this system does not perform well when tested with audio clips of people with stutters or vocal damages. The expression of emotions are a very personalized element for a particular person. Some people show intense emotions while the others express emotions in a more reserved manner.

9.3 Recommendations for Future Developments

The ability of the system to adapt over the time will enhance the overall performance by presenting a more personalized behavior. As of future improvements for this system the addition of semantic information of the speech can be considered to improve the overall accuracy of determination of the interest level and affective state of the user. Even further by using a multi modal approach by including a visual channel, the system will be able to determine the attention level of the user when both listening and speaking.

Another limitation of this system is not being properly able to identify the entities that can either be considered as a noun or a verb. The ambiguous nature of such instances can cause the system to not detect the relevant entities of the utterance. Overall this system can be considered as a starting point for the robots to extract the entities with emotional significance of the user which will fulfill their quest to be longterm companions of the users. Since human perception of the entities change over the time and new entities are introduced through course of the time it is important for the system to update accordingly.

One of the areas for future research is to find a method to coordinate empathetic behavior in a multi-party conversation. In such a case their may be several parties showing empathy and there should be a way for a robot to be a part of such a conversation. Showing empathy in multi-party conversation will differ than in a private conversation due to the fact that the governing social norms are different hence the robot's behavior should be modified accordingly. In order to facilitate real time processing a dynamic sampling method is used which would further make possible to identify vocal features such as stress and sighs.

LIST OF PUBLICATIONS

---

1. S. D. Wickramaratne and A. B. P. Jayasekara, Attention level approximation of a conversation in human-robot vocal interaction using prosodic features of speech, in Engineering Research Conference (MERCon), 2017 Moratuwa. IEEE, 2017, pp. 4045. (*Published*)

2. S.D.Wickramaratne and A. G. B. P. Jayasekara. " Emotionally Significant Memory Determination of a User Experience for a Domestic Service Robot". (*In Preperation*)

REFERENCES

[1] "World population ageing 2015," ST/ESA/SER.A/390, Population Division, Department of Economic and Social Affairs, United Nations, 2015.

[2] D. O. Johnson, R. H. Cuijpers, J. F. Juola, E. Torta, M. Simonov, A. Frisiello, M. Bazzani, W. Yan, C. Weber, S. Wermter, *et al.*, "Socially assistive robots: a comprehensive approach to extending independent living," *International journal of social robotics*, vol. 6, no. 2, pp. 195–211, 2014.

[3] E. Broadbent, R. Stafford, and B. MacDonald, "Acceptance of healthcare robots for the older population: review and future directions," *International Journal of Social Robotics*, vol. 1, no. 4, p. 319, 2009.

[4] D. Feil-Seifer and M. J. Matarić, "Socially assistive robotics," *IEEE Robotics & Automation Magazine*, vol. 18, no. 1, pp. 24–31, 2011.

[5] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, pp. 22–35, 2015.

[6] T. Haidegger, M. Barreto, P. Gonçalves, M. K. Habib, S. K. V. Ragavan, H. Li, A. Vaccarella, R. Perrone, and E. Prestes, "Applied ontologies and standards for service robots," *Robotics and Autonomous Systems*, vol. 61, no. 11, pp. 1215–1223, 2013.

[7] H. Takeda, N. Kobayashi, Y. Matsubara, and T. Nishida, "Towards ubiquitous human-robot interaction," in *Working Notes for IJCAI-97 Workshop on Intelligent Multimodal Systems*, pp. 1–8, 1997.

[8] T. Darrell and A. Pentland, "Space-time gestures," in *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on*, pp. 335–340, IEEE, 1993.

[9] E. Prassler, G. Lawitzky, A. Stopp, G. Grunwald, M. Hägele, R. Dillmann, and I. Iossifidis, *Advances in Human-Robot Interaction*, vol. 14. Springer Science & Business Media, 2004.

[10] Z. Obrenovic and D. Starcevic, "Modeling multimodal human-computer interaction," *Computer*, vol. 37, no. 9, pp. 65–72, 2004.

[11] N. Roy, G. Baltus, D. Fox, F. Gemperle, J. Goetz, T. Hirsch, D. Margaritis, M. Montemerlo, J. Pineau, J. Schulte, *et al.*, "Towards personal service robots for the elderly," in *Workshop on Interactive Robots and Entertainment (WIRE 2000)*, vol. 25, p. 184, 2000.

[12] C. L. Sidner and C. Lee, "Engagement rules for human-robot collaborative interactions," in *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, vol. 4, pp. 3957–3962, IEEE, 2003.

[13] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[14] R. Plutchik, "The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[15] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[16] J. A. Russell, "Core affect and the psychological construction of emotion.," *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[17] H. Lövheim, "A new three-dimensional model for emotions and monoamine neurotransmitters," *Medical hypotheses*, vol. 78, no. 2, pp. 341–348, 2012.

[18] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE transactions on affective computing*, vol. 5, no. 2, pp. 101–111, 2014.

[19] E. Shouse, "Feeling, emotion, affect," *M/c journal*, vol. 8, no. 6, p. 26, 2005.

[20] P. A. Thoits, "The sociology of emotions," *Annual review of sociology*, vol. 15, no. 1, pp. 317–342, 1989.

[21] R. J. Dolan, "Emotion, cognition, and behavior," *science*, vol. 298, no. 5596, pp. 1191–1194, 2002.

[22] R. Plutchik, "The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[23] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[24] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[25] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[26] R. Khosrowabadi, H. C. Quek, A. Wahab, and K. K. Ang, "Eeg-based emotion recognition using self-organizing map for boundary detection," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 4242–4245, IEEE, 2010.

[27] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[28] J. A. Russell, J.-A. Bachorowski, and J.-M. Fernández-Dols, "Facial and vocal expressions of emotion," *Annual review of psychology*, vol. 54, no. 1, pp. 329–349, 2003.

[29] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[30] R. Cowie and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, pp. 1989–1992, IEEE, 1996.

[31] Y. Li, L. Chao, Y. Liu, W. Bao, and J. Tao, "From simulated speech to natural speech, what are the robust features for emotion recognition?," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pp. 368–373, IEEE, 2015.

[32] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural computing & applications*, vol. 9, no. 4, pp. 290–296, 2000.

[33] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Fuzzy emotion recognition in natural speech dialogue," in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, pp. 317–322, IEEE, 2005.

[34] A. A. Razak, R. Komiya, M. Izani, and Z. Abidin, "Comparison between fuzzy and nn method for speech emotion recognition," in *Information Technology and Applications, 2005. ICITA 2005. Third International Conference on*, vol. 1, pp. 297–302, IEEE, 2005.

[35] R. Tokuhisa, K. Inui, and Y. Matsumoto, "Emotion classification using massive examples extracted from the web," in *Proceedings of the 22nd Inter-*

national Conference on Computational Linguistics-Volume 1, pp. 881–888, Association for Computational Linguistics, 2008.

[36] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1, pp. I–577, IEEE, 2004.

[37] B. Schuller, M. Lang, and G. Rigoll, "Robust acoustic speech emotion recognition by ensembles of classifiers," *Fortschritte der Akustik*, vol. 31, no. 1, p. 329, 2005.

[38] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 119–155, 2003.

[39] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[40] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," *Handbook of face recognition*, pp. 247–275, 2005.

[41] C. E. Osgood, W. H. May, and M. S. Miron, *Cross-cultural universals of affective meaning*, vol. 1. University of Illinois Press, 1975.

[42] C. E. Osgood and O. Tzeng, *Language, meaning, and culture: The selected papers of CE Osgood.* Praeger Publishers, 1990.

[43] J. T. Hancock, C. Landrigan, and C. Silver, "Expressing emotion in text-based communication," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 929–932, ACM, 2007.

[44] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 206–218, 2011.

[45] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[46] J. Bhaskar, K. Sruthi, and P. Nedungadi, "Hybrid approach for emotion classification of audio conversation based on text and speech mining," *Procedia Computer Science*, vol. 46, pp. 635–643, 2015.

[47] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*, pp. 205–211, ACM, 2004.

[48] J. Weizenbaum, "Elizaa computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[49] K. M. Colby, "Modeling a paranoid mind," *Behavioral and Brain Sciences*, vol. 4, no. 4, pp. 515–534, 1981.

[50] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Hierarchical neural network generative models for movie dialogues," *CoRR, abs/1507.04808*, 2015.

[51] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "A communication robot in a shopping mall," *IEEE Transactions on Robotics*, vol. 26, no. 5, pp. 897–913, 2010.

[52] H. Wang, F. Zhang, X. Fan, and X. Lu, "A practical service robot system for greeting guests," in *Control Conference (CCC), 2012 31st Chinese*, pp. 4997–5001, IEEE, 2012.

[53] D. Bohus, C. W. Saw, and E. Horvitz, "Directions robot: in-the-wild experiences and lessons learned," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 637–644, International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[54] D. Jurafsky, "Speech and language processing: An introduction to natural language processing," *Computational linguistics, and speech recognition*, 2000.

[55] J. Cassell, "Embodied conversational interface agents," *Communications of the ACM*, vol. 43, no. 4, pp. 70–78, 2000.

[56] J. R. Wilson, "Towards an affective robot capable of being a long-term companion," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pp. 754–759, IEEE, 2015.

[57] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[58] S. Haq and P. J. Jackson, "Multimodal emotion recognition," *Machine audition: principles, algorithms and systems*, pp. 398–423, 2010.

[59] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech.," in *Interspeech*, vol. 5, pp. 1517–1520, 2005.

[60] R. M. Agrigoroaie and A. Tapus, "Developing a healthcare robot with personalized behaviors and social skills for the elderly," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pp. 589–590, IEEE Press, 2016.

[61] R. V. Ibanez, M. U. Keysermann, and P. A. Vargas, "Emotional memories in autonomous robots," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pp. 405–410, IEEE, 2014.

[62] E. A. Kensinger, "Negative emotion enhances memory accuracy: Behavioral and neuroimaging evidence," *Current Directions in Psychological Science*, vol. 16, no. 4, pp. 213–218, 2007.

[63] C. Kuhbandner and R. Pekrun, "Joint effects of emotion and color on memory.," *Emotion*, vol. 13, no. 3, p. 375, 2013.

[64] M. A. V. J. Muthugala and A. G. B. P. Jayasekara, "Mirob: An intelligent service robot that learns from interactive discussions while handling uncertain information in user instructions," in *Moratuwa Engineering Research Conference (MERCon), 2016*, pp. 397–402, IEEE, 2016.