Affect Recognition with Vocal Intonation for Human-Robot Interaction

by

Yuma Tsuboi

A thesis submitted in conformity with the requirements
for the degree of Master of Engineering

Department of Mechanical and Industrial Engineering
University of Toronto

Affect Recognition with Voice Intonation for Human-Robot Interaction

Yuma Tsuboi

Master of Engineering

Department of Mechanical and Industrial Engineering
University of Toronto

2016

Abstract

The advancements of its sub-disciplinary fields has propelled robotics to become one of the most promising technology to solve the issues in variety of domains and it is expected that there will be more robots in the daily lives of people in the near future. Accordingly, a great effort is being made to increase the quality of human robot interaction (HRI), particularly in the area of socially assistive robots and its ability to engage in a natural bidirectional communication with people. This thesis presents the development of vocal intonation modal for multimodal affect recognition for the prediction of the affect state of the person whom a robot is interacting. Various features for classification of speaker's affect state has been investigated. A total of 51 vocal intonation features were determined to be effective when classified with a model that was trained using the multi-layer perceptron neural network learning algorithm. This classification model was then used to develop a real-time vocal intonation affect recognition system.

Acknowledgments

I am very grateful for the supervision of Professor Goldie Nejat. Her guidance has been invaluable in completing my research and having it published in international conferences and journals.

I would like to thank Alex Hong, Nolan Lunscher, Tianhao Hu and Leo Woiceshyn for their cooperation and assistance in my research. I could not have finished my thesis without their contributions.

A special thanks goes to Alex Hong for putting up with me while we worked very closely this past year. His admirable enthusiasm and work ethics has been a great inspiration in surmounting the challenges faced during this research.

I am thankful for the encouragement from Veronica Marin to partake in this research for the Master of Engineering project. Her advice has introduced me to an enriching experience with many challenges and opportunities.

Thank you to all graduate students of the Autonomous Systems and Biomechatronics Laboratory for their assistance and support throughout this research.

Most importantly, I would like to express my deepest gratitude to my mother, father, Luffy and all of my friends, for their continuous understanding and support. Their love and blessings are what made this thesis possible.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1
Introduction

1 Introduction

1.1 Motivation

Gone are the days where machines were built to simply carry on the tasks we as humans want completed. The field of robotics has seen progress in its sub-disciplines and its potential to further serve the needs of people has been established to the public. As robots become more relevant in our lives, improving the quality of interaction between human and robots, or human-robot interaction (HRI), has become a subject of great interest. With higher quality of HRI, a more natural interaction can take place between a human and robots, leading to appreciation for their companionship and acceptance of the coexistence of robots in our lives. The elderly is a particular age group who will benefit from such natural social interaction. As the population age, the elderly will require increased assistance with daily needs and find themselves with less support, as they will significantly outnumber the rest of the population. In addition to the materialistic needs such as food and shelter, the elderly will require care and companionship, which is not always feasible for family and friends to provide. A robot that can interact with high quality HRI can provide the necessary companionship for the elderly.

1.2 Socially Assistive Robots

Socially Assistive Robotics (SAR) is an area in robotics that focus on providing social interaction to people [1]. Its application includes assisting people to achieve positive progress in convalescence, rehabilitation and learning by having a close and effective interaction [1]. Socially assistive robots are often designed to interact with people in a human-like way, for example through speech, facial expressions and body language [2]. The execution of this bi-directional communication requires the robot to understand and respond to many complex contexts. One way to stimulate natural HRI is for the robot to understand the affect state of the person whom it is interacting [3]. The interpretation of the other's affect state becomes particularly important in interactions with elderly, including those diagnosed with dementia, because they may have difficulty speaking but instead communicates using non-verbal communication, such as facial expression, body language and vocal intonation [4]. The social

robot must respond appropriately with an emotion of its own for the bidirectional communication to take place. A natural approach to accomplish this is for the robot to possess an affect state of its own. This leads to the need of emotion modeling which can transform dynamically according to the prediction of the affect state of the person whom it is interacting.

1.3 Vocal Intonation Modal of Affect Recognition

All socially living species possess the ability to vocalize affect state. The affect state of human species in particular, with the development of speech, has become audible in speech [5]. Research of the vocal expression of affect state has active for short period of time and the correlation between affect state and vocal intonation cues is still unclear [6]. However, certain affect states has clear influence on particular physiological states from which effect on vocal intonation can be predicted [7]. For example, when one is in a state of anger, fear or joy, the speech becomes loud, fast and enunciated with strong high frequency energy. This is because the sympathetic nervous system becomes aroused, which leads to increased heart rate and blood pressure and results in drier mouth and occasional muscle tremors. When one is bored or sad, the speech becomes slow, low-pitched speech with little high-frequency energy. This is because the parasympathetic nervous system becomes aroused, which leads to decreased heart rate and blood pressure and salivation increases [8]. These physiological contributions to the change in vocal intonation is same across all human species, which suggests that the correlation between one's vocal intonation and affect state is independent of their culture and languages [9]. This makes vocal intonation a reliable and frequently used modal of affect communication that cannot be ignored in creating an affect recognition system.

1.4 Challenges

Robots with social intelligence can provide effective response to the speaker with the consideration for their affect state in addition to the verbal contents. This leads to a natural bi-directional interaction, which promotes the acceptance of social robots and encourage deeper engagements. But the challenge is to design such robot that can interpret the complex social behaviors of people. The representation of affect state is also challenging because of the difference in the interpretation of communication cues between individuals and the many models that exist in psychology to describe the complex human affect. Additional complication arises

from the accuracy of the sensors for communication cues that carry affect information. By interpreting all modes of non-verbal communication, the robot can make the informed conclusion on the speaker's affect state. In particular, the affect recognition from the vocal intonation modal is a difficult problem because of the need for speaker independency. Human performance on speaker independent affect recognition was surveyed and only 60 percent of the participants were able to correctly predict the expressed emotion of unknown people [10], which indicates that it is not an easy task even for human.

1.5 Problem Statement and Thesis Objective

The overall goal of this research is to investigate the effectiveness of multimodal affect recognition and emotional modeling by a social robot in increasing the quality of HRI. The multimodal system is composed of body language and vocal intonation modal, as displayed in the overall flowchart of each component of this research in figure 1.

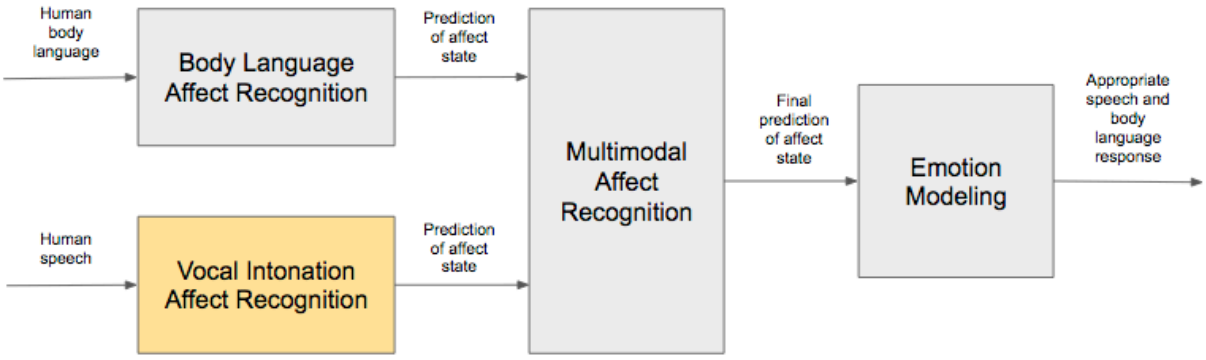

Figure 1: Flow chart of the overall research of affect state recognition and emotion modelling for improved HRI

This overall research is broken down into four sub-components, which are being worked on by four students from the Autonomous Systems and Biomechatronics (ASB) Lab at the University of Toronto, and for this reason the scope of this thesis has been set to focus on one of these components. This thesis focuses specifically on the research of the effectiveness of the vocal intonation modal in affect recognition in HRI, marked yellow in figure 1. Hence, the objective of this thesis is to analyze the accuracy of affect recognition using solely the vocal intonation of the utterance and to develop a real-time system to predict the affect state of the speaker from their

speech. The intention is to integrate the resultant real-time vocal intonation affect recognition system to the multimodal affect recognition architecture, which is outside the scope of this thesis.

Chapter 2
Literature Review

2 Literature Review

For robots to coexist and have a meaningful social interaction with people, it must be able to interpret the affect state of the people whom it is interacting. As discussed in Chapter 1, the affect recognition from vocal intonation is a challenging task but also necessary in achieving a natural bidirectional conversation [10]. This chapter reviews the literature on the state of vocal intonation affect recognition, models used to represent affect state from psychology, feature extraction methods from speech utterances and classification techniques for the prediction of affect states that correspond to the given utterance.

2.1 Vocal Intonation Affect Recognition

The psychology of emotion has been propelled notably due to the research on the facial expression modal by Darwin [11]. In his work, Darwin has emphasized the correlation of facial and vocal expression of emotions in both humans and animals [12]. The inference of emotion from voice is particularly interesting with the case of human because of our unique ability to predominantly communicate intentions through speech. However, emotion recognition from this powerful method of communication has been largely neglected perhaps due to the invisibility of sound and the technological challenges in recording it for analysis [11]. In 1986, Scherer has evaluated the progress and the future direction of research in this field by testing variety of acoustic features such as loudness, tempo and variety of fundamental frequency (perceptually heard by human as pitch) related parameters. It was concluded with some signs of correlation between the emotional state of the speaker and their vocal muscle action patterns and the anticipation of advancements in digital speech signal analysis addressing the difficulty of emotion recognition from voice. Scherer also emphasized the importance of evaluating the human performance of emotion recognition from vocal intonation as a bench mark of this technology and reported in 1995 that people on average can accurately predict the emotion associated with the speech by roughly 60% [5]. This is far better than correctly predicting by chance, which indicates that there exist vocal cues perceived by humans to make informed

predictions. Scherer has compiled a list of audible cues that portray emotional state, which is displayed in Table 1. Although these cues are qualitative rather than quantitative, the development list was a major step in identification of the features that can be used to infer emotions. Modern research on affect recognition from vocal intonation typically uses the approach of quantifying various vocal intonation features and performing classification to infer the affect state of the speaker.

Table 1: Affect states and their audible cues during speech [5]

Affect state	Audible Cues
Sad	Down pitch contour, Low pitch level, Slow tempo, round envelope
Excited	Large pitch variation, small amplitude variation, fast tempo, sharp envelope, high pitch level
Anger	Small pitch variation, up pitch contour, high pitch level, fast tempo
Boredom	Small pitch variation, low pitch level, slow tempo, round envelope
Happy	Small amplitude variation, Large pitch variation, Fast tempo, Sharp envelope
Disgust	Small pitch variation, round envelope, slow tempo, up pitch contour [13]
Interest	Fast rate of speech and wider range of frequency [14], extreme pitch variation and rising contour [15]
Surprise	Fast tempo, high pitch level pitch contour up, sharp envelope, large pitch variation
Fear	Pitch contour up, fast sequence, high pitch level, round envelope, small pitch variation

2.2 Affect Models

There are two types of mainstream models used in psychology to represent human affect states: categorical and dimensional models, which are shown in figure 2. Categorical model maps perceived affect state to one of the finite number of discrete emotions. It is widely accepted that there are 6 basic and distinct emotions with clear boundaries, which are happy, sad, angry, fearful, surprised and disgusted [16]. The advantage of this model is that one can distinguish

each category of emotion clearly. The downside is that it is difficult to agree on a suitable set of categories that can address the complexity of the affect states of a person and that more categories leads to longer computation time. On the other hand, the dimensional model represents affect state by using one or multiple vectors to create a continuous spectrum of emotions. The strength of this model is that it can address the complex mixed emotions not included in the finite categorical model [17]. The dimensional models can also better cope with the emotional state caused by ambiguous stimuli than categorical models [17].

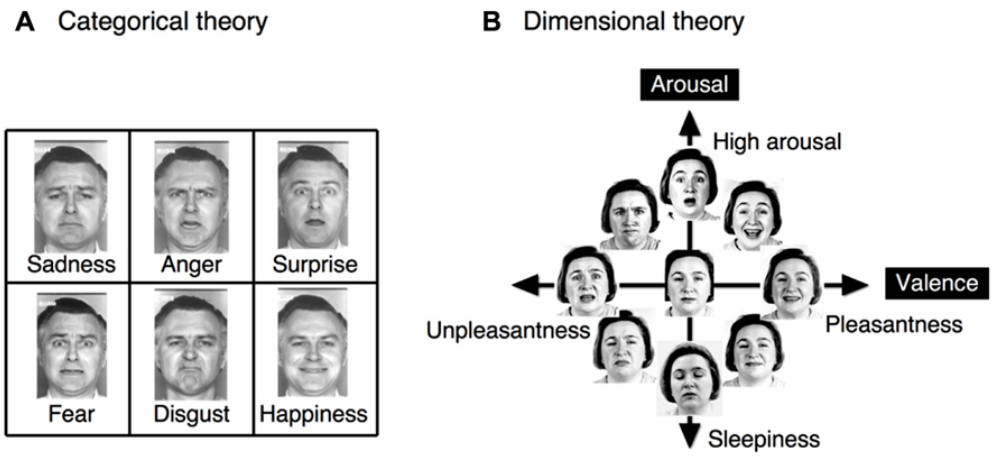

Figure 2: Categorical model (left) and dimensional model (right) [16]

Multidimensional model has been popularized by Plutchik and Russel who has postulated a circumplex space model based on two fundamental dimensions of valence and arousal [11]. Hereby called the circumplex model, shown in figure 3, can represent each emotion as a linear combination of the two variables and allows one to graphically illustrate the similarities and differences between neighboring emotions [18]. The range of valence and arousal value are both {-2, -1, 0, +1, +2}. Table 2 shows the valence and arousal values for different affect states.

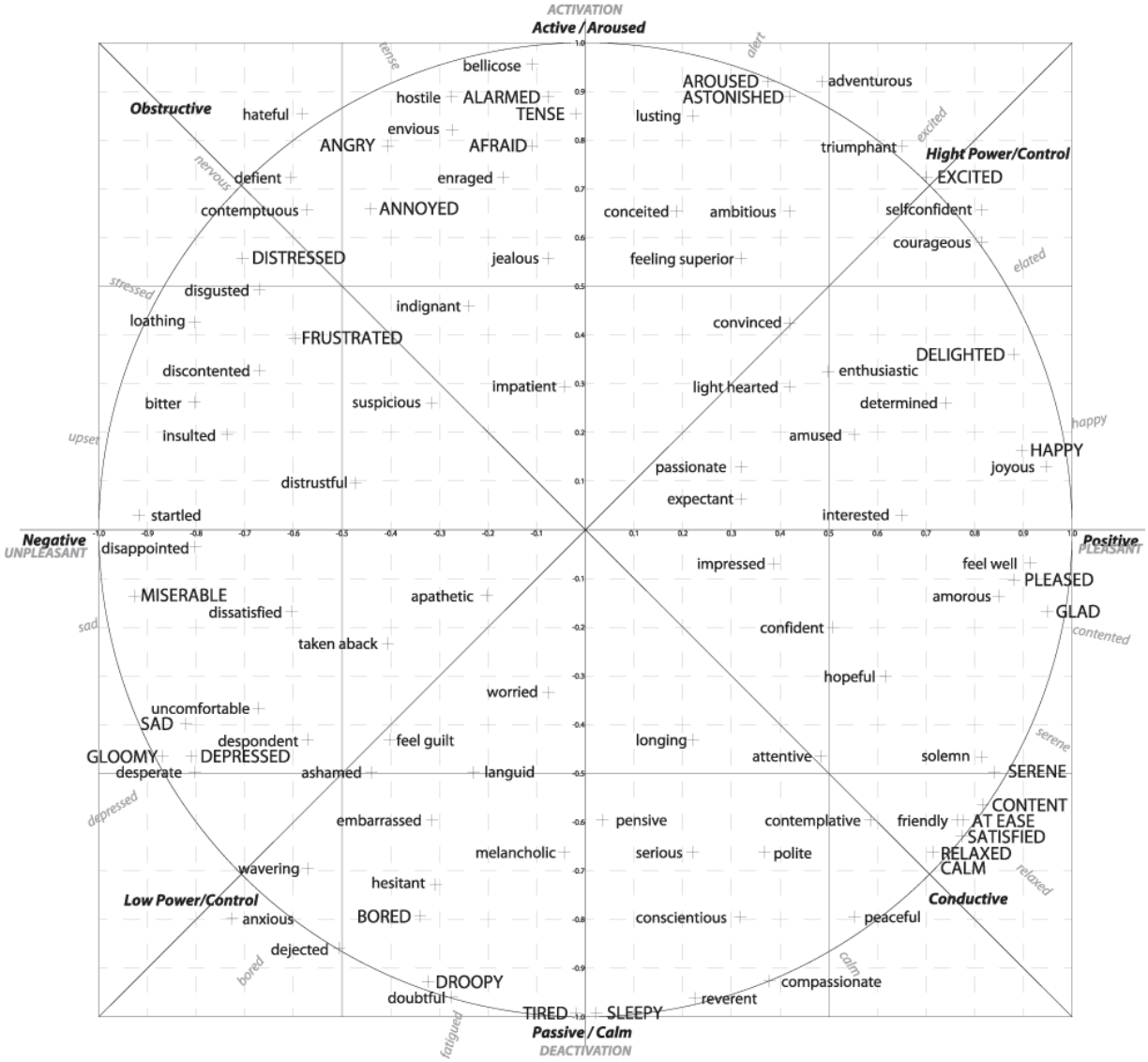

Figure 3: Two dimensional circumplex model of affect [19]

Table 2: Valence and arousal values for each affective state

Affect state	Valence	Arousal
Sad	-2	-1
Excited	+2	+2
Anger	-1	+2
Bored	-1	-2

Happy	+2	0
Disgust	-2	+1
Interest	+1	0
Surprise	0	+2
Fear	0	+2

2.3 Database of Emotional Speech

The affect models provide systematic method of representing affect states. However, the interpretation of utterances to a specific affect state is still a subjective problem. In order to objectively map a certain emotional utterance to the corresponding affective state with consistency, there needs to be a set of examples that encompass all the possible situations. Online databases of recorded emotional utterances are a collection of utterances that are labelled with the affect state that was adopted. Vocal intonation features can be extracted from these utterances to build a database of training data for affect classification. Sections 2.3.1, 2.3.2, 2.3.3 presents the assessment of few of the publically available online emotional utterance databases.

2.3.1 EmoDB

Database consists of records from 5 male actors and 5 female actresses, expressing 6 different emotions. They speak in German and this sample collection was conducted by Technical University of Berlin from 1997 to 1999 [20]. This database contains a reasonable number of speakers speaking the same verbal content to offer both generalization over the target group and comparability across affect state and speakers [21]. The utterances are performed with acted emotions due to the rarity of expression of real emotions in the real world and the ethical issue of its recording if it were to be captured [21].

2.3.2 RML

Database that contains 720 audiovisual emotional expression samples collected by the Ryerson Multimedia Lab. One of the six emotions are expressed in each video clip: anger, disgust, fear happiness, sadness and surprise. The video samples were collected from eight human subjects, speaking six different languages (English, Mandarin, Urdu, Punjabi, Persian, Italian). This is

advantageous in testing for language independency in the affect recognition. The samples were recorded at a sampling rate of 22050Hz using a single channel 16-bit digitization and a frame rate of 30 fps. Each video clip has a length of about 3-6 seconds [22].

2.3.3 SEMAINE

This database was created by the Queen's University of Belfast. It contains emotional speech in 4 emotional states: anger, sadness, happiness and sensible. The readers are 40 volunteers (20 females, 20 males) aged between 18 to 69 years. The subjects read 5 passages of 7-8 sentences written in an appropriate emotional tone and content for each emotional state. Each passage was read with strong expression of the corresponding emotional state [23] [24]. This database contains utterances by speakers of age above 55, who can represent the elderly population.

2.4 Commercial Social Robot as an Affect Recognition Platform

The database of emotional speech is a representation of the speaker utterances a robot will hear during social interaction. From this, the vocal intonation affect recognition can be developed but it must be installed on a robot to test its real-time performance. Conveniently, there are programmable robots that are commercially available and their high level of autonomy makes them suitable as a platform to install the affect recognition system to observe the simulation of a social interaction. A humanoid robot is a popular choice of design for this application with people because it can interact with the environment in a similar and familiar way as the human do, which makes it understandable by humans [25]. The following subsections discusses three commercially available social robots that are around or under the price of $10,000 CAD that the vocal and the multimodal affect recognition can be installed.

2.4.1 Pepper

The Pepper robot, developed by Aldebaran Robotics and SoftBank, is a four-feet tall humanoid robot designed to live with humans. As a social robot, it has features like to converse, react to emotions, move and live autonomously with people [26]. It is equipped with a variety of sensors including four microphones, 2 HD cameras, a 3-D depth sensors, 2 gyroscopes, touch sensors, two sonars to operate in the same environment as humans. It can express emotions through its blinking lights on the eyes, ears and mouth and it is capable of displaying body gestures. Pepper is already installed with an internal AI that allows it to converse and understand, but it can run

custom programs written in Python and C++. A Pepper SDK for Android Studio has been released and there are open source projects host on the Community website of Aldebaran, which makes Pepper very attractive to developers [27]. Pepper was released to consumers in Japan in June 2015 for $1600 USD and it is planned to release in the United States in 2016. The release date for Canada is still not known, but the capability of Pepper and its price is a good display of the current state of the robotics industry.

Figure 4: Pepper robot

2.4.2 Poppy

Poppy is a three feet tall, 3-D printed humanoid robot. The face is a small display and its skeleton is visible from the outside. There is a wide angle USB camera on the head and sensors embedded in its motors that allow for gestures and bipedal locomotion. Poppy is an open source robot with various software and 3-D printing configurations that can not only help developers, but also aid scientists who want to do more research on bipedal locomotion, or physical or social interaction between robots and humans [25]. It is priced at around 8500 USD.

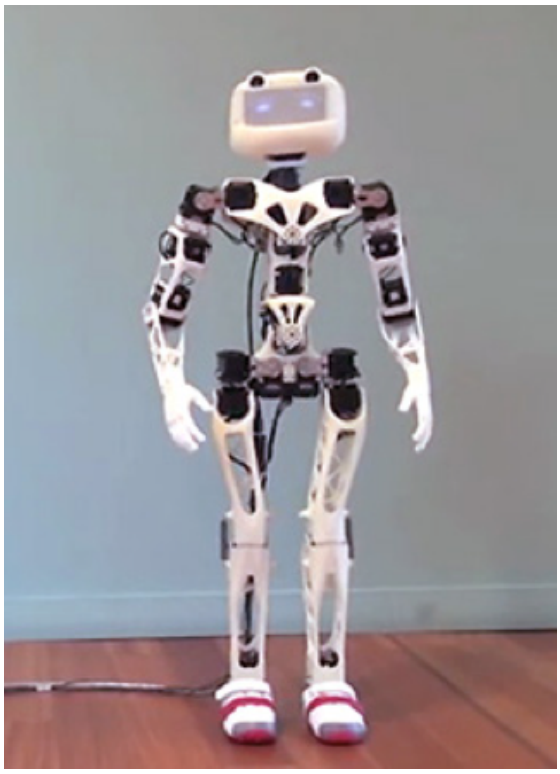

Figure 5: Poppy robot

2.4.3 NAO

The NAO robot, also by Aldebaran and SoftBank, is a programmable humanoid robot equipped with two cameras, four directional microphones, sonar, two IR emitters and receivers, one inertial board, nine tactile sensors and eight pressure sensors. Standing at just 58 cm, NAO is the first humanoid robot developed by the company and there are already 9000 units sold and is in its 5th version [28]. Aldebaran has an established cross-platform build tools, core communication library and well designed user community page that contain blogs, tutorials and important announcements that attracts developers to tailor NAO to many different applications. These applications include taking part in a dance choreography, playing in the Robocup 2015 (robotics world cup for soccer), interacting with children with autism at a specialized center and aplatform for developing emotional applications at SoftBank Robotics [28]. The NAP robot is available for purchase for $7990 USD.

Figure 6: NAO robot

2.5 Feature Selection for Affect Classification

The affect recognition system that is to be installed on a social robot solves the classification problem of the identifications of the relevant affect states as defined by a specific affect model, using techniques from machine learning and data mining. That is, in order to predict the affect state associated with speech utterances, one must extract the vocal intonation features that carry affect information from the given audio signal. The selection of the features is a challenging task since the correlations between the different features and the affect states are still inconclusive. Nevertheless, there is a general consensus on the feature set that can lead to high successful affect classification rate, such as pitch, energy, duration and mel-frequency cepstral coefficients (MFCC). This section discusses the two tools that can be used to extract a set of features from recorded utterance to be used for affect classification.

2.5.1 Nemesysco QA5 SDK

The QA5 is a software development kit (SDK) for emotion detection and measurement, published by Nemesysco Ltd [29]. Its application includes integration to call centers technology to monitor anger and stress during conversations for quality assurance [29]. Given a voice

segment data, the QA5 SDK produces a set of 17 emotional levels of the speaker, which can be used as features for affect classification: energy, content, upset, angry, stress, embarrassment, intensive thinking, imagination activity, hesitation, uncertainty, excitement, concentration level, SAF, extreme state, atmosphere, brain power and EmoCog ratio. The problem is that these features are values processed from fundamental vocal intonation features and therefore not intuitive to understand. The mapping between the vocal intonation features and the emotional levels is not disclosed by Nemesysco for confidentiality. According to their public patent, the QA5 technology is based on the layered voice analysis (LVA) technology, which uses the plateau and thorns statistics from the given audio signal to infer the emotional values [30]. Thorns are a local maximum in a triplet of digitized sound sample that are above an arbitrary threshold, while plateaus are detected when the samples in a triplet have absolute maximum amplitude deviation [31]. Unfortunately, the information available to the public about the algorithm behind LVA is insufficient to verify its validity. Nevertheless, established organizations such as the United States Special Operations Command and Department for International Journal of Speech utilizes the Nemesysco program. Thus, it is worthwhile to investigate its performance in affect recognition through experimentation.

2.5.2 Matlab Audio Analysis Library

The Matlab Audio Analysis Library is a public library for audio analysis available on the Matlab File Exchange. It is the code offered alongside the book “Introduction to Audio Analysis, A MATLAB Approach” by Theodoros Giannakopoulos and Aggelos Pikrakis. The library includes variety of audio analysis tasks including general audio handling (e.g. I/O, recording), audio processing and feature extraction [32]. The transparency of the library allows development of the vocal intonation affect recognition application feasible with more control than with QA5 SDK. The library contains function for the extraction of the following vocal intonation features: zero-crossing rate, energy, spectral entropy, spectral roll off, spectral centroid spread, spectral flux, mel-frequency cepstral coefficients, fundamental frequency and chroma vector. These features, in particular the zero crossing rate, fundamental frequency and mel-frequency cepstral coefficients, are widely accepted in many related works to have strong correlation with the change in affect states [33] [34]. Thus, it can be anticipated that these feature will be effective in the prediction of the affect state of speaker and is worthwhile to investigate.

2.6 Affect Classification

The vocal intonation features extracted from recorded utterance contains underlying patterns that are reflective of the speaker's affect information. The accurate recognition of these patterns will lead to affect recognition and it can be achieved using classification techniques from machine learning and data mining. Instead of developing our own code for classification, an open source data mining tool was used to assess the most suitable classification algorithm. Weka, an open source data mining software created by The University of Waikato, is a collection of classification algorithms that simplify data mining tasks [35]. The ability to preprocess the training data and to perform cross validation with the resultant classification model makes it feasible to experiment and determine the most effective classifier for affect recognition. In addition to the graphical user interface, this software also offers script commands that makes it well suited for real time applications.

Chapter 3
Emotional Utterance Recording

3 Emotional Utterance Recording

As described in 2.1, the general procedure of developing a real-time vocal intonation affect recognition system begins with the offline training of the classification model that can predict the corresponding affect state of a given stream of audio data in a real-time manner. The classification model is created using supervised learning algorithm, which requires training data, or a collection of pairings between a feature vector and the corresponding label. These are training examples that the classification model will use to determine the label when a new unlabeled example is inputted in real-time application. In this context, the feature vector and the label are the vocal intonation features and the corresponding affect state, respectively, and it is important that the collection is representative of all possible affect state that a speaker may experience in real life. To construct a database of labeled feature vectors, it is necessary to obtain emotional utterance recordings where the affect state experienced by the speaker is known. Section 3.1 discusses the approach taken with SEMAINE database of emotional speech that is publically available. Section 3.2 discusses the collection of emotional utterances from human samples in a controlled environment.

3.1 SEMAINE Database

As presented in section 2.3.3, the SEMAINE database contains emotional speakers who are aged above 55 years old, the age group considered to be an elderly. However, there only three affect states that are relevant to our research: angry, sad and happy. Nevertheless, the resultant database would useful for testing the performance of vocal intonation affect classification. Angry, sad and happy speeches by two subjects, aged 58 and 57 at the time of recording, were examined. The 6 speeches were segmented into clips that contains the whole utterance, or sentences, to eliminate the silence between phrases. As a result, 50 emotional utterance audio clips were created for each emotion. This would guarantee at least a total of 150 training data when feature extraction is performed, because each training data is based on about 2 second segments of audio data. One problem with this resultant database of training data is that there is no corresponding information

about the speaker's body gesture during the speech. Although this is not important for the scope of the affect recognition from vocal intonation, it becomes relevant during the training for the multimodal affect recognition, which requires both synchronized feature vectors of affect classifications from vocal intonation and body language modals.

3.2 Custom collection of emotional utterances

The utterance database created from the SEMAINE database is sufficient for the experimentation with vocal intonation modal of affect recognition, vocal intonation training data with corresponding body language training data is required for the overall multimodal affect classification. Inspired by the approaches taken to create databases of emotional speech, our very own database of emotional speech was created.

3.2.1 Subjects

The emotional speeches were recorded in the ASB Lab. Three subjects from the ASB Lab were asked to read the passage from the Snow White fairytale, which is presented in the Appendix. The passage was read nine times by each subject with the following affect states: happy, sad, angry, disgust, fear, surprise, excited, bored and interested. The first six affect states are the universal emotions as outlined in [16] and the latter three are those deemed to be important in determining the engagement level of the elderly during interaction. These affect states were expressed to the best of their ability with the audible cues outlined in table 1. Although the subjects have no background in emotion acting, the systematic guidance of table 1 and the number of subjects will introduce variance in the speaker's acted affect state, which is important in representing wider range of target. In total, 18 audio clips of emotional speech were collected with the speakers engaging in expressive body languages for the multimodal training in the future.

3.2.2 Speech Recording

The Voice Tracker II Array Microphone by Acoustic Magic was used to record its noise cancelling and speaker tracking technology. Using Audacity, the free open source audio recorder and editor, the speeches were sampled at 11025 Hz with 16-bit depth. The audio data was exported in a .wav format for compatibility with the Nemesysco's QA5 SDK and the Matlab

Audio Analysis Library discussed in Chapters 4 and 5. The circumplex model is utilized for this research for its ability to represent different affect states in a continuous space. The exported .wav files were titled with the information of the associated affect states so that it can be converted to the respective valence and arousal values of {-2, -1, 0, +1, +2}, following feature extraction.

3.2.3 Physical Setup

The physical setup during the collection of the emotional speech should be same as during the real-time application and also should be representative of a natural social interaction with people. To emulate this scenario, the NAO robot is chosen as the platform for the affect recognition system and is placed in front of the speaker as a placeholder. The NAO robot were chosen over the other two social robots described in section 2.4 for its availability, price range and proven reliability indicated by the positive reviews. Since NAO is 58 cm in height, it is placed on a table in its standing stance during the speech recording to be close to the eyelevel of the speaker. The Voice Tracker microphone is placed in front of NAO but on the edge of the table so that it would not obstruct NAO from moving during the real-time application. The distance between the speaker and NAO should be 0.75 – 1.3 m, range of distance determined to be comfortable for people during socialization [36]. 0.75m was chosen so that the speaker can be close as possible to the microphone. The Snow White script was placed in front of NAO so that the subject can still make an eye contact. The computer that receives the audio signal from the microphone is hidden from the view of the speaker. Figures 7 and 8 shows the physical setup during the collection of the emotional speech.

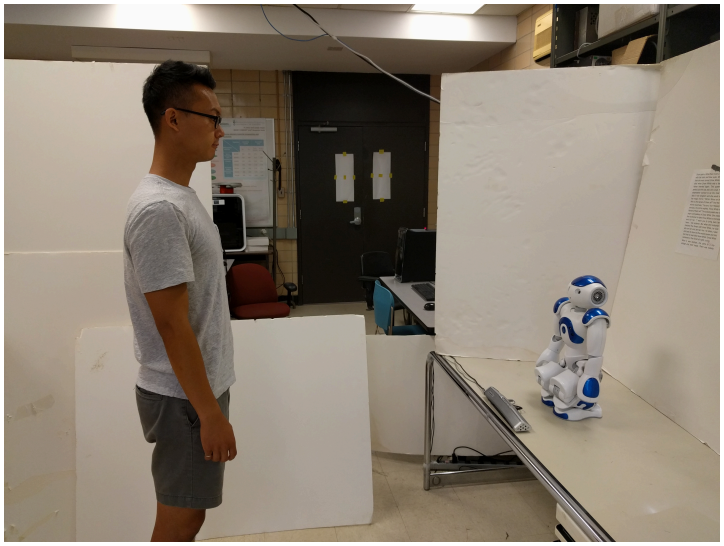

Figure 7: Side view of the physical setup. The microphone is located at the feet of NAO

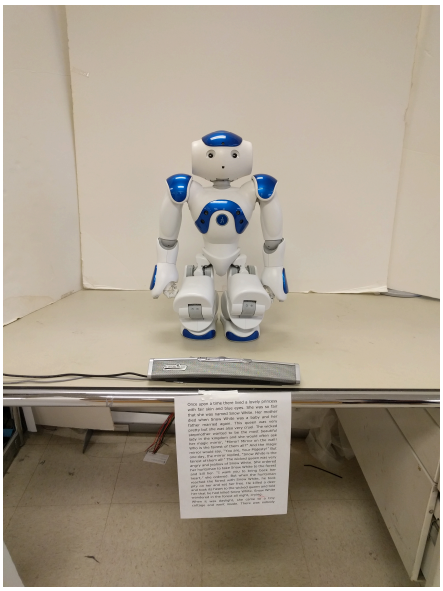

Figure 8: The view from the person interacting with NAO. The Snow White excerpt is located under the microphone

3.2.4 Noise Reduction

Once the speech is recorded, a low pass filter is then applied to further reduce noise. With the Nemesysco's QA5 application, the SDK offers a configuration for expected background noise level and the setting recommended by the user manual is used. With the Matlab application, a moving average filter is applied for its straightforward implementation and its ability to reduce random noise while retaining a sharp step response [37]. It operates by averaging the surrounding samples to produce each point in the output. Below is its equation:

$$\boldsymbol{y[i]} = \frac{\mathbf{1}}{\boldsymbol{M}} \sum_{\boldsymbol{j=0}}^{\boldsymbol{M-1}} \boldsymbol{x[i+j]} \tag{1}$$

where $x[i]$ is the input signal with noise, $y[i]$ is the filtered output signal and M is the number of samples in the average, in this case 30. With this collection of filtered emotional speeches, the vocal intonation feature extraction can take place.

Chapter 4
Classification with QA5 SDK

4 QA5 SDK by Nemesysco

4.1 Nemesysco Features

As described in section 2.3.1, the QA5 SDK allows extraction of 17 features from a given audio data. These features are non-traditional vocal intonation features that was derived using the LVA algorithm, which is unfortunately undisclosed to the public. The features are summarized in table 3.

Table 3: Nemesysco QA5 features [29]

Feature	Description	Ranges
Energy (E)	Perhaps the most indicative parameter QA5 exposes. Conversation energy indicates if the speaker is sad or tired at low values (below 5), comfortable (5-9) or highly energetic (above 9). Values at the range of 0 to 1 may also indicate boredom.	Range is 0 to 50. Normally expected values are between 3 to 9.
Content (C)	Indicates how pleased or happy the tested party is, but due to the psychological nature of different conversation modes, in times may appear in an argument as well.	0 to 30, normal value is 0

Upset (Ups)	Indicates how unpleased or sad the tested party is.	0 to 30, normal value is 0
Angry (A)	Indicates how angry the tested party is. Please note that the system may accidently detect anger in different conversation moods.	0 to 30, normal value is 0
Stress (S)	Indicates how nervous the tested party is.	0 to 30, normal value is 0
Embarrassment (Em)	Indicates how uncomfortable the tested party is.	0 to 30, normal value is 0
Intensive Thinking (IT)	Indicates the tested party is thinking intensively while speaking.	0 to 30, normal value is below 3
Imagination Activity (IA)	Indicates that the tested party is either recalling information from his memory or visualizing something.	0 to 30, normal value is 0
Hesitation (H)	Indicates how comfortable the tested party was when making the statement. Below 14, the subject is more comfortable than normal;	0 to 30, normal range is 14-17

	above 17, the subject is regretting what he is saying.	
Uncertainty (Unc)	Indicates how certain or uncertain the tested party is about what was said. Below 15, the subject is more certain; above 15, the subject is more uncertain.	0 to 30, normal value is **15**
Excitement (Exc)	Indicates how positively or negatively excited the tested party is. Below 15, excitement is negative; above 15, indicates higher excitement or anger.	0 to 30, normal value is **15**
Concentration Level (CL)	Indicates how concentrated the tested party is.	0 to 30, normal value is 0
SAF	Arousal factor, indicating deep and profound interest in the conversation topic (positive or negative!).	0 to 30, normal value is 0
Extreme State (ES)	A value indicating how extreme the overall emotional activity is according to a unique logic pre-programmed in the QA5 Core.	0 to 30, normal value is 0
Atmosphere (At)	A general value indicating the emotional atmosphere during	(-100) to 100, normal range is (–10) to 10.

	the conversation. Negative values normally indicate general uncomfortable feelings; positive values indicate positive feelings.	
Brain Power (BP)	Overall summary of both emotional and cognitive processes in the brain. Used mainly for research purposes, but may also be useful assisting through research for your own needs and alert definitions.	0 – 6000 (both theoretical) Normal ranges are between 700 and 1000
EmoCog Ratio (ECR)	Indicates rationality of the subject, i.e., to what extent the subject is talking based on emotions or logic. Above 100, the subject is more emotional; below 100, the subject is more logical.	1 to 5000 (both theoretical) Normal ranges are between 50 to 150

4.2 Offline Training

In order to create a real-time affect recognition program, a classification model that can predict the affect state from a given utterance must be trained offline. This section describes the offline training process, which includes the collection of training data, experimentation for the optimal configuration of the QA5 SDK and the comparison between the classifiers for highest performing model.

4.2.1 Building the Training Data

A database of training example is necessary to generate the classification model that can classify the affect state in real time for given segment of utterance. Section 4.2.1.1 describes the feature extraction performed on the custom database of emotional speech. Section 4.2.1.2 describes the feature normalization on the resultant feature vector and 4.2.1.3 describes the its labeling with the corresponding affect states to create training data. Finally, section 4.2.1.4 describes the compilation of each training example into a database, or a training data, and its formatting for the Weka program to receive.

4.2.1.1 Feature Extraction

At the time of development of the classification system with QA5 SDK, there were only two subjects available for the construction of the emotional speech database described in section 3.2. Hence, there was a total of 18 audio clips, which were segmented into 2 second audio data to represent utterance, which is the average length it lasts [38]. Feature extraction program was developed with the QA5 SDK based on the code written by Bijan Shahriari from the ASB Lab, which is written in C# and runs in C++ [39]. The program generated a total of 363 feature vectors of length 17 each.

4.2.1.2 Labeling feature vector

When feature vectors are labelled with the corresponding class, it becomes a training example that exhibit predictive relationships between the utterance and associated affect state. The extracted feature vectors were written to a .csv file along with the title of the .wav file of the utterance, which contains the affect state information of the speaker. A python script was written to obtain that information and convert into valence and arousal values, as shown in table 2. These values were appended to the feature vector. Figure 9 shows the format of a single training example for valence and arousal.

E	C	Ups	S	Em	IT	IA	H	Unc	Exc	CL	SAF	ES	At	BP	ECR	V

(a)

E	C	Ups	S	Em	IT	IA	H	Unc	Exc	CL	SAF	ES	At	BP	ECR	A

(b)

Figure 9: Training example for (a) valence and (b) arousal

4.2.1.3 Conversion to Weka Format

The training examples were compiled into a list to create the training data in .csv format. In order to process the training data with Weka, it must be converted into an .arff format, which is described in [68]. A python script was written to convert the training data in .csv format into .arff format with the appropriate headings that declare the feature names and its data type. Two training data in .arff format was generated, for valence and arousal.

4.2.2 Classification Optimization

Two classification models were trained with the Weka GUI, for valence and arousal. The classification problem here is to predict the valence and the arousal values, which has the range of {-2, -1, 0, +1, +2}. The performance of each classification model was measured using 10-fold cross validation with the training data. This classification rate was improved using two methods, by experimenting with the configurations of the QA5 SDK for the feature extraction stage and by selecting the appropriate learning algorithms in Weka.

4.2.2.1 QA5 Configuration Optimization

Two parameters for QA5 configuration was experimented to obtain the training data that can be used to train the optimal classification model. The first parameter determines the time length of the utterance segments for analysis. QA5 can be configured to analyze streaming voice data in lengths of 1, 2 and 3 seconds. The user manual recommends that 2-second-long voice segments will produce the most stable result [29], which is consistent with the average length of utterances [38]. The option for 3 seconds was not tested because it would decrease the size of the training

data. The second parameter is the calibration type, which is the SDK configuration that determines the calibration procedure and the analysis mode. There are two types of calibration procedure, full and short. In a full calibration process, first 10 segments are used for calibration, while in a short calibration process, only the first segment is used. Unfortunately, the different analysis modes are simply enumerated 1-3 and their details are not disclosed by Nemesysco Ltd. According to [29], its optimal setting depends on the hardware and codec of the devices used and therefore requires experimentation to find it. The combination of the two calibration procedures and the three analysis modes results in six calibration types that needs to be investigated. Table 4 summarizes these calibration type that determines the SDK's protocol for feature extraction.

Table 4: Calibration type parameter for QA5 SDK [29]

Calibration Type	Calibration Procedure	Analysis Mode
0	Full	1
1	Short	1
2	Full	2
3	Short	2
4	Full	3
5	Short	3

4.2.2.2 Learning algorithm Optimization

Using Weka, the following classifiers were experimented to determine the most suitable for affect recognition: Naïve Bayes, multinomial logistic regression (logistic regression), multi-layer perceptron neural networks (MLP), k-nearest neighbors (KNN), random forest decision tree and classification via regression. These were selected to investigate the different types of learning. Naïve Bayes is a type of probabilistic learning with feature independence assumption. Multinomial logistic regression is a linear regression analysis that does not assume independence between the features but instead assumes collinearity [40]. MLP is a feedforward neural networks algorithm that can detect flexible decision boundaries. KNN is a type of lazy learning that does not require training a predictive model in advance [41]. Random forest decision tree is a combination of tree predictors that is robust with respect to noise in the data [42].

Classification via regression, or model tree, is a decision tree with linear regression functions at the leaves [43]. A 10-fold cross validation was used to compare the performances of each classifier.

4.3 Results

4.3.1 Optimal QA5 Configuration

In order to determine the calibration type and segment length that produce the highest classification rate, all combinations were tested by extracting the features and training a model using the classification via regression classifier. Table 5 shows the 10-fold cross validation performances of the models derived by all combinations. In conclusion, the segment length of 2 seconds and calibration type of 5 produced the overall highest classification rates for both valence and arousal.

Table 5: Cross validation result of models trained with all combinations of segment lengths and calibration types

Segment length/calibration type	Valence Classification Rate	Arousal Classification Rate
1/0	50.2%	55.3%
1/1	49.3%	56.4%
1/2	50.3%	55.0%
1/3	49.2%	55.7%
1/4	50.0%	55.4%
1/5	49.6%	56.0%
2/0	53.0%	50.2%
2/1	53.9%	56.9%
2/2	53.0%	57.9%

2/3	53.9%	56.2%
2/4	52.9%	58.8%
2/5	54.1%	56.9%

4.3.2 Learning Algorithm for Optimal Classifier

For the classification model for valence, the random forest decision trees obtained the highest classification rate of 55.6%. But because the model requires noticeable amount of time to classify a test data, it is inapplicable for real-time applications. Hence, the highest performing model was trained by classification via regression based model, with 54.2%. As for arousal, the logistic regression classifier produced the highest performing model with 59.0%.

Table 6: Learning techniques tested for valence and arousal

Classifier	Valence Classification Rate	Arousal Classification Rate
Naïve Bayes	44.9%	50.0%
Logistic	49.5%	59.0%
Multi-layer Perceptron Neural Networks	49.8%	55.7%
K-Nearest Neighbors	46.0%	48.9%
Random Forest Decision Tree	55.6%	58.5%
Classification via regression	54.2%	57.2%

4.4 Discussion

Through experimentation of the different combinations of QA5 configurations, it was determined that the optimal classification model was trained with segment length of 2 seconds, type 5 calibration. Training based on classification via regression classifier and logistic regression produced the optimal performing model for valence and arousal, respectively. However, these models produced cross validation results of 54.2% and 59.0%, which are relatively low and will

not add sophistication to the overall multimodal system. This result could be improved by using feature normalization and experimentation with feature selection, although a significant increase in the recognition rate cannot be expected. More importantly, the lack of transparency of the QA5 SDK, such as the undisclosed details of the feature derivation and analysis mode is an obstacle in further optimizing the affect recognition performance. For these reasons, the second method using the Matlab Audio Analysis Library is investigated.

Chapter 5
Classification with Matlab Audio Analysis Library

5 Classification with Matlab Audio Analysis Library

In this chapter the program for feature extraction and affect classification using the Matlab Audio Analysis Library is developed. This is a component in the overall development of the real-time affect recognition system that is analogous to that developed using QA5 SDK discussed in Chapter 4. As such, the presentation of the development process will follow very much the same way. Section 5.1 describes the feature selection process with the definition, motivation and computation of each. Section 5.2 describes the offline training using the features selected and the optimization process of the classification models. Section 5.3 presents the result of the performances of the classification models and section 5.4 discusses the results and its implications for the overall system.

5.1 Feature Selection

With the Matlab Audio Analysis Library, a program was developed that use a total of 51 features to classify the valence and arousal of utterances. These features can be categorized into two groups: 35 fundamental features, which measures unique characteristics in speech signals, and 16 optimization features, which were experimentally shown to increase the classification rate. The fundamental features are: zero crossing rate (ZCR), spectral centroid spread, spectral flux, 18 Mel-frequency cepstral coefficients (MFCCs), fundamental frequency, and chroma vector of 12 values. These features can be further divided into three groups, prosodic, spectral and tonality. The prosodic features describe the way speech was spoken (e.g. punctuation, volume), while the spectral features describe the vocal tract configuration, articulation, which are physical characteristics of the human vocal system. Tonality features are the strength of each pitch classes in the speech. By combining these three types of features, each will be represented in the classification, thus making the final prediction more sophisticated. The optimization features are: energy, spectral entropy, spectral roll off, and 14 linear prediction filter coefficients (LPCs). The optimization features introduce redundancy in the measured characteristics of the utterance but used due to their unique definition and the higher classification rate produced during

optimization. The definition and motivation of the usage of all the features are described in the subsection below.

5.1.1 Fundamental Features

5.1.1.1 Zero crossing rate (ZCR)

ZCR is the rate of sign-changes along a signal and it indicates the proportion of the speech signal that are voiced and unvoiced [44]. It is defined as

$$\boldsymbol{ZCR_n} = \sum_{\boldsymbol{m}=-\infty}^{\infty} |\boldsymbol{sgn}[\boldsymbol{x}(\boldsymbol{m})] - \boldsymbol{sgn}[\boldsymbol{x}(\boldsymbol{m}-\boldsymbol{1})]|\boldsymbol{w}(\boldsymbol{n}-\boldsymbol{m}) \tag{2}$$

where

$$sgn[x(n)] = \begin{cases} 1, x(n) \geq 0 \\ -1, x(n) < 0 \end{cases}$$

and

$$w(n) = \begin{cases} \frac{1}{2N}, 0 \leq n \leq N-1 \\ 0, otherwise \end{cases}$$

and N is the number of samples in the frame of interest [45]. ZCR is high for unvoiced because the signal is close to zero and low for voiced because it is far from zero. It is a measure of the speed and the volume of the speech by the speaker. This is associated with the physiological changes in speaker's respiration, which carries emotional state information [46] [47]. The ZCR is computed with Matlab by comparing the pair of adjacent points and counting the number of times the sign change occurs.

5.1.1.2 Spectral centroid spread (SCS)

Spectral centroid spread is the variance of the spectral centroid, or the measure of the bandwidth of the spectrum. It is the weighted mean of the frequencies in the signal, with frequency magnitudes as weights and it is defined as:

$$\boldsymbol{SC(n)} = \frac{\sum_{n=0}^{N-1}(\boldsymbol{f(n)} - \boldsymbol{SC(n)})^2 \boldsymbol{x(n)}}{\sum_{n=0}^{N-1} \boldsymbol{x(n)}} \quad (3)$$

Where n is the bin number, $x(n)$ is the weighted frequency value or the magnitude of n, $f(n)$ is the center frequency of n and $SC(n)$ is the spectral centroid defined as:

$$\boldsymbol{SC(n)} = \frac{\sum_{n=0}^{N-1} \boldsymbol{f(n)x(n)}}{\sum_{n=0}^{N-1} \boldsymbol{x(n)}} \quad (4)$$

[48] [49]. Spectral centroid spread can characterize the physiological changes in articulation, which carries emotional information [46] [50].

5.1.1.3 Spectral flux (SF)

Spectral flux is the measure of how quickly the power spectrum of a signal is changing [51]. It is defined as:

$$SF(t) = |X(t, w) - X(t-1, w)| \quad (5)$$

Where $X(t, w)$is the power spectrum of frame t. In general, speech is quasi-stationary and slow-changing signal over a short time frame. Hence, the spectral flux can be used to determine speech from other rapidly changing sounds [52]. Specifically, it has been shown to be useful for distinguishing between fear (valence=0, arousal=2) and anger (valence=-1, arousal=2) [53]. The spectral flux is computed in Matlab as the sum of the squared distances between the normalized spectrum of the current and the previous frame.

5.1.1.4 Mel-frequency cepstral coefficients (MFCC)

MFCC is a representation of the Mel-frequency cepstrum, a short-term power spectrum of a sound based on a linear cosine transform of a log power spectrum on a nonlinear Mel-scale of

frequency. It is reflective of human hearing perception, which cannot perceive frequencies over 1000 Hz [54]. There are 7 steps in computing the MFCCs, as outlined in [54] and [55]:

1. Pre-emphasis

   The signal is passed through a filter to emphasize the higher frequencies, in order to increase the high frequency energy.

   $$\boldsymbol{y[n] = x[n] - ax[n-1]} \tag{6}$$

   where $a = 0.95$ makes 95% of any one sample is derived from the previous sample.

2. Framing

   The 2 second segment is further broken down into frames of 20ms.

3. Hamming windowing

   The hamming window is applied to each sample, which minimizes the maximum side lobe [56] [57]. The Hamming window coefficients are computed using:

   $$\boldsymbol{W[n] = 0.54 - 0.46\cos\left(2\pi\frac{n}{N}\right), 0 \le n \le N} \tag{7}$$

   where N is the number of samples in each frame [58]. The coefficients are applied to the signal simply by:

   $$\boldsymbol{Y[n] = X[n] \times W[n]} \tag{8}$$

4. Fast Fourier transform

   The samples in each frame is converted from time domain into frequency domain using the equation:

   $$\boldsymbol{Y(w) = FFT[H(t) * X(t)] = H(w) * X(w)} \tag{9}$$

   where $H(w)$ is the vocal tract impulse response.

5. Mel-filter bank processing

The frequency ranges in the FFT spectrum is very wide and voice signal does not follow the linear scale. 27 overlapping Mel-scale triangular filters, shown in figure 10, is applied to compute amount of energy in each filterbank.

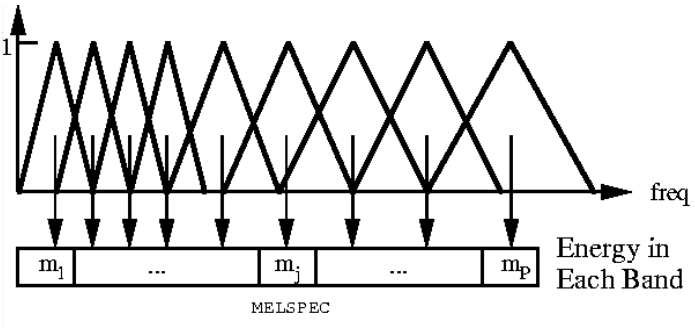

Figure 10: Mel-scale filter bank [59]

6. Take logarithm

   Logarithm is applied to each of energies, which produces 27 log filterbank energies.

7. Discrete cosine transform (DCT)

   Convert the log Mel-spectrum into time domain by applying DCT. This generates 27 MFCCs and the first 18 are used as features for affect recognition, as described in section 5.2.3.

MFCC characterizes the vocal tract configuration, which carries emotional information [46]. The typical number of MFCCs used is 13-20, as any higher or lower did not help with recognition [60].

5.1.1.5 Fundamental frequency (F_0)

Fundamental frequency is the lowest frequency of a periodic waveform. Autocorrelation method was used to compute the fundamental frequency of each segment. The autocorrelation function at time position n_0 is

$$\boldsymbol{C_x(d) = \sum_{n=n_0}^{n_0+W} x(n)x(n-d)} \tag{10}$$

where W is the length of the summation window and d is the time lag. For periodic signals, $C_x(d)$ will start from its maximum value and the fundamental period, T_0, is the first position of d where $C_x(d) = C_x(0)$ [61]. Figure 11 shows a periodic signal and its corresponding autocorrelation function labelled with the location of the period.

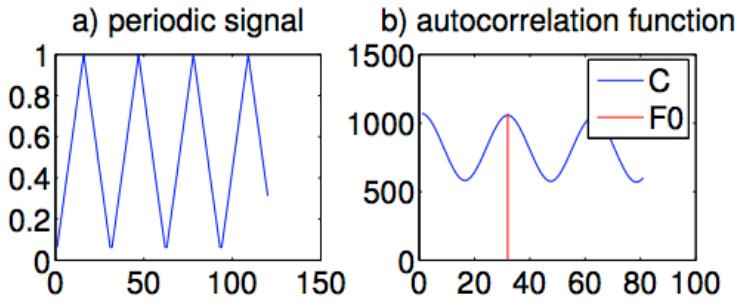

Figure 11: A periodic signal and its autocorrelation function. The red line indicates the location of the first peak where the the fundamental period ends [61]

The fundamental frequency, f_0, is then computed by:

$$\boldsymbol{f_0 = \frac{1}{T_0}} \tag{11}$$

Fundamental frequency, or pitch, is directly related to emotions in speech, including anger, joy, fear, disgust and sad [62].

5.1.1.6 Chroma vector (CV)

Chroma vector is a pitch distribution feature that describe tonality, measuring the energy of each of the 12 pitch classes of the equal-tempered scale within an analysis frame. Therefore, chroma vector describes the 12 semitones composition of the audio signal and represent the tonality content of the speech [63]. The 12-dimensional chroma vector vector is computed by taking sthe sum of the log-frequency magnitude spectrum across octaves:

$$\boldsymbol{C(b) = \sum_{z=0}^{Z-1} |X_{lf}(b + z\beta)|} \tag{12}$$

where $X_{lf}$is the log frequency spectrum, z is the integer octave index, Z is thenumber of octaves, b is the integer pitch class index and β is the number of bins per octave [64] [65]. Inspired by its

use in musical applications to perceive emotions, a study showed that it is effective in detecting neutral and sad emotions [64], which would translate to negative valence.

5.1.2 Optimization Features

5.1.2.1 Energy (E)

Energy is the size of signal, or the area under the signal curve and it can indicate the proportion of voice segments that are voiced and unvoiced [44]. It is defined as:

$$\boldsymbol{E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2} \tag{13}$$

which is simply the area under the curve for the frame of interest [45]. Hence, energy is high for voiced and low for unvoiced. High energy indicates voiced Energy is an alternative measure of determining the speed and the volume of the speech by the speaker. This is associated with the physiological changes in speaker's respiration, which carries emotional state information [46] [47]. The extraction of energy is coded in Matlab as the area under the signal curve.

5.1.2.2 Spectral entropy (SE)

Spectral entropy is the measure of spectral distribution. It quantifies a degree of randomness of spectral probability density represented by normalized frequency components of the spectrum [46]. It is defined as:

$$\boldsymbol{H(|Y(w,t)|^2) = -\sum_{w=1}^{\Omega} P(|Y(w,t)|^2)\log\left(P(|Y(w,t)|^2)\right)} \tag{14}$$

where

$$P(|Y(w,t)|^2) = \frac{|Y(w,t)|^2}{\sum_{w=1}^{\Omega}|Y(w,t)|^2}$$

is the probability of the frequency band w for the magnitude spectrum of frame t [66]. Entropy is high for unvoiced segments and low for voiced segments, thus it is useful for detecting the voiced and unvoiced components of the speech [67]. This is associated with the physiological changes in speaker's respiration, which carries emotional state information [46] [47].

5.1.2.3 Spectral roll off (SRO)

Spectral roll off is Nth percentile of the power spectral distribution of the audio signal, where N is usually chosen to be 85-95%. It is computed as the frequency index R below which the majority of the spectral energy is concentrated:

$$\sum_{k=1}^{R} \boldsymbol{X_t(k)^2} \leq \boldsymbol{N} \sum_{k=1}^{\frac{M}{2}} \boldsymbol{X_t(k)^2} \tag{15}$$

where $X_t(k)$is the magnitude spectrum of the signal frame computed by an M-point fast Fourier transform [68]. It was shown to be useful for distinguishing between fear (valence=0, arousal=2) and disgust (valence=-2, arousal=1) [69]. This two dimensional range is slightly different than the range represented by spectral flux, and its inclusion resulted in higher classification rated during the optimization.

5.1.2.4 Linear predictive coding (LPC) coefficients

LPCs are coefficients from approximating current signal as the linear combination of the previous samples:

$$\boldsymbol{s(n)} = \sum_{k=1}^{p} \boldsymbol{\alpha_k s(n-k)} \tag{16}$$

where p is the number of previous samples and α_k is the LPC coefficients [70]. LPC calculates the power spectrum and the formant frequencies, frequency where resonant peaks occur [71]. LPC is one characterization of the shape of the speaker's vocal tract which carries emotional information [46], similarly to MFCC. 14 LPCs were empirically found to be the ideal number of coefficients to use, as described in section 5.2.3.

5.2 Offline Training

The valence and arousal of the speaker are classified using the 51 vocal intonation features. In order to develop the classification models for valence and arousal, a similar approach with the QA5 SDK offline training was conducted. Feature vectors were extracted from utterances and

were labelled with the corresponding valence and arousal values to produce training examples for each classification. Then, a learning algorithm was selected for each based on the nature of the data and cross validation performance of the trained models.

5.2.1 Building Training Data

In order to generate the classification models that can classify the valence and arousal in real time for a given segment of utterance, a database of training examples for each was created. Section 5.2.1.1 describes the feature extraction performed on the custom database of emotional speech. Section 5.2.1.2 describes the feature normalization on the resultant feature vector and 5.2.1.3 describes the its labeling with the corresponding affect states to create a training example. Finally, section 5.2.1.4 describes the compilation of all training examples into two training data for valence and arousal, and its formatting for the Weka program to process.

5.2.1.1 Feature Extraction

Each of the 27 emotional utterance clips from the extracted from the custom emotional speech database described in section 3.2 were segmented into 2 second audio data to represent utterance. Feature extraction was performed using a program developed in Matlab with the Matlab Audio Analysis Library to generate 994 feature vectors of length 51. This resulted in a total 994 feature vectors.

5.2.1.2 Feature Normalization

As a preprocessing of the data, each features extracted in the previous step was scaled to [0, 1] range using the min-max normalization method:

$$\boldsymbol{x'} = \frac{\boldsymbol{x} - \mathbf{min}\,(\boldsymbol{x})}{\mathbf{max}(\boldsymbol{x}) - \mathbf{min}\,(\boldsymbol{x})} \tag{17}$$

where x is the original value of the feature and the x' is the normalized value. Normalization of the data speeds up convergence during training and helps avoid getting stuck in local optima [72]. It also introduces standardization in the data with different units and scales, which is important because many data mining techniques rely on Euclidean distance which is likely to be governed by features with wider range of values [73]. Multi-layer perceptron neural networks is a data mining technique that is sensitive to feature scaling [74].

5.2.1.3 Labeling Feature Vectors

After normalization, the feature vectors were written to a .csv file along with the title of the .wav file that it was extracted from, which contains the affect state information of the speaker. A python script was used to use that information and convert into valence and arousal values, as shown in table 2. These values were appended to the feature vector. Figure 12 shows the format of a single training example for valence and arousal.

ZCR	SCS	SF	SRO	F_0	CV 1	...	CV12	MFCC 1	...	MFCC 18	E	SE	LPC 1	...	LPC 14	V

(a)

ZCR	SCS	SF	SRO	F_0	CV 1	...	CV12	MFCC 1	...	MFCC 18	E	SE	LPC 1	...	LPC 14	A

(b)

Figure 12: Training example format for (a) valence and (b) arousal

5.2.1.4 Conversion to Weka Format

The resultant training examples were compiled into a list to create the training data in .csv format. In order to process the training data with Weka, it must be converted into an .arff format, which is described in [68]. A python script was used to convert the training data in .csv format into .arff format with the appropriate headings that declare the feature names and its data type. Two training data for valence and arousal was generated in the .arff format.

5.2.2 Classification Optimization

Two classification models were trained using the Weka GUI, for valence and arousal. The classification problem here is to predict the valence and the arousal values, which has the range of {-2, -1, 0, +1, +2}. The performances of each classification models were measured using 10-

fold cross validation with the training data. This classification rate was improved using two methods, by experimenting with the configurations of the QA5 SDK for the feature extraction stage described in section 5.2.2.1 and by selecting the appropriate learning algorithms in Weka described in section 5.2.2.2.

5.2.2.1 Learning Algorithm Selection Optimization

Using Weka, various learning algorithms were tested with the valence and arousal training data and evaluated using cross validation. Analogously with the learning algorithm selection with QA5, we compared the cross validation classification rate of the following classifiers: Naive Bayes, logistic regression, MLP, KNN, random forest decision tree, classification via regression. Parameters associated with respective learning algorithm were adjusted empirically to achieve the highest classification rate.

5.2.2.2 Feature Selection Optimization

With the most effective classifier, the contribution of each feature in successful classification was examined. This was performed experimentally by training classification models using different combinations of features and comparing the cross validation result. A feature vector of 35 fundamental features was first tested and the optimizing features were added subsequently to see its effectiveness in increasing the classification rate. This procedure is also used to determine the optimal number of coefficients for MFCC and LPC.

5.3 Classification Results

The valence and arousal training data each comprised of 51 vocal features were used to train a model with different learning algorithms and were evaluated on classification correctness. This result is summarized in table 7. The highest classification rate was achieved with MLP for both valence and arousal classification, which were 79.4% and 84.4%, respectively. This could be due to the classifier's strength in the ability to detect non-linear boundaries in the data, which can be

anticipated due to the large number of features [75]. Another reason could be due to the amount of noise that we can expect to be in the data, which neural networks can better handle than random forest decision, which obtained the second highest classification rate [76].

Table 7: Cross validation results for models trained with different classifiers

Classifier	Valence Classification Rate	Arousal Classification Rate
Naïve Bayes	54.5%	70.9%
Logistic	71.4%	75.7%
Multi-layer Perceptron Neural Networks	79.4%	84.4%
Simple Logistic	70.3%	82.6%
Random Forest Decision Tree	71.0%	78.6%
Classification via regression	69.4%	72.6%

With the MLP classifier, the effectiveness of the feature selection was tested experimentally. Firstly, the number of MFCC coefficient to optimize the classification rate was determined. Table 8 describes the process of this optimization for the number of MFCC, where we used the average of valence and arousal for comparison. The fundamental features with 18 MFCCs produced the highest classification rate of 71.5% and 79.7% for valence and arousal, respectively.

Table 8: Optimization process for the number of MFCCs

Feature vector	Valence (%)	Arousal (%)	Average (%)
Fundamental features (without MFCCs)	59.1	59.8	59.5
13 MFCCs	67.2	75.5	71.4

14 MFCCs	68.7	76.1	72.4
15 MFCCs	68.8	78.2	73.5
16 MFCCs	70.8	78.3	74.6
17 MFCCs	70.3	78.9	74.6
18 MFCCs	71.5	79.7	75.6
19 MFCCs	70.6	78	74.3
20 MFCCs	72.4	78.5	75.5

Next, the optimization features were added incrementally to the fundamental feature vector, as summarized in table 9. We looked at the average classification rate between the valence and arousal for comparison between the different feature vectors. The optimization features were added to the fundamental features in the order of: energy, spectral entropy and LPCs. The number of LPCs is also experimented from 10-16 (see the LPC description). In conclusion, empirically the classification rate increased when energy, spectral entropy and 14 LPCs were added to the fundamental feature vector. The optimal classification rate achieved was 79.4% and 84.4% for valence and arousal, respectively.

Table 9: Optimization process for the optimization features

Feature vector	Valence (%)	Arousal (%)	Average (%)
Fundamental features	72.4	78.5	75.45
Energy	72.9	79.5	76.2
Spectral entropy	73.6	79.4	76.5
10 LPCs	77.2	82.9	80.05
11 LPCs	77.4	82.9	80.15

12 LPCs	76	80.4	78.2
13 LPCs	77.6	82.2	79.9
14 LPCs	79.4	84.4	81.9
15 LPCs	79.5	82.9	81.2
16 LPCs	78	84.4	81.2

As a result, the 51 features determined to be useful are: ZCR, spectral centroid spread, spectral flux, 18 MFCCs, fundamental frequency, chroma vector, energy, spectral entropy, spectral roll off, and 14 LPCs. Feature vector using these 51 features are used for classifying the affect state of the speaker from their voice.

5.4 Discussion

Through experimentation using the Weka GUI, it was determined that the optimal performing affect classification models can be trained with MLP, which produced a cross validation result of 79.5% and 84.4%, for valence and arousal, respectively. After, the feature selection was systematically altered to test the classification performance of the model derived using different subsets of features. The optimal number of MFCCs was tested first, followed by the effect of each of the optimization features. In conclusion, 18 MFCCs produced the highest classification performance and the optimization features of energy, spectral entropy and 14 LPCs were found to further increase the classification rate. Due to the large number of possible combinations of the features and learning algorithms, not all were tested. However, the systematic inclusion of the optimization features showed that they were effective in successful affect classification, although the speech characteristic that they measure suggested that it would only introduce redundancy to the measurements by the fundamental features. With the optimized classification model for valence and arousal, the real-time affect recognition system can be developed.

Chapter 6
Real Time Affect Recognition

6 Real Time Affect Recognition

The classification models for valence and arousal generated with the MLP classifier is used to predict the affect state of a speaker in real time.

6.1 System Architecture

The architecture for this system, which is shown in Figure 13, is comprised of 3 modules: voice recognition (VR), feature extraction (FE), affect classification (AC).

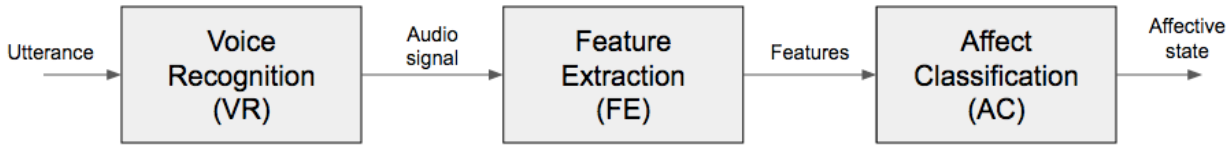

Figure 13: Architecture for real time affect recognition from vocal intonation

6.1.1 Voice Recognition

The VR module captures the utterance exactly the same way as how the audio for the training data was obtained. Hence, the utterance is sampled at 11025 Hz with 16-bit depth with the Voice Tracker II Array Microphone and then passed through a low pass filter to eliminate noise. In real time, the VR module is constantly sampling the sound and and for every 2 seconds of audio signal, it is processed and sent to the FE module.

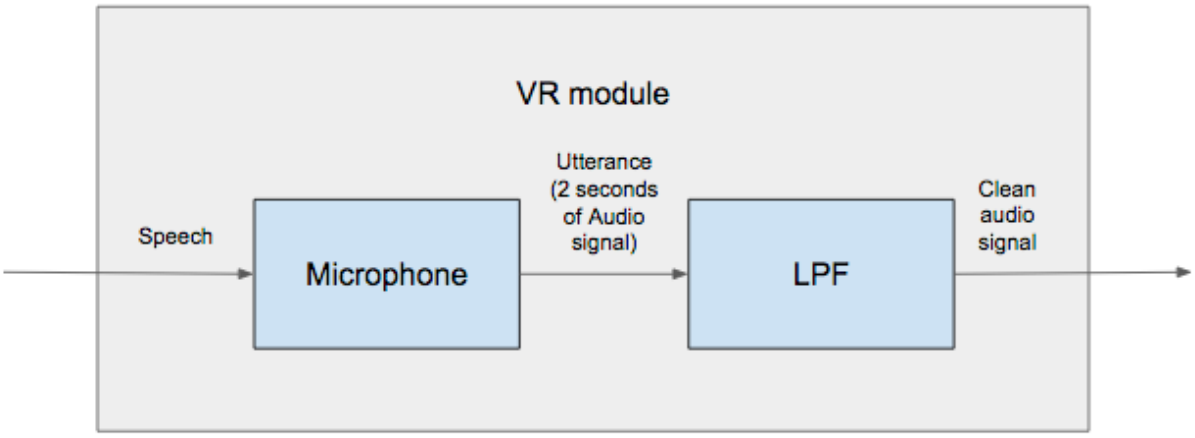

Figure 14: Voice recognition module

6.1.2 Feature Extraction

The FE module receives the audio signal from VR and extracts the vocal feature vector. A Matlab program was developed with the Matlab Audio Analysis Library to extract the 51 features, discussed in section 5.1, from the audio signal segments. Once the feature vector is obtained, it is processed into a specific format for unlabeled instance and written to an. arff file and send to the AC module to be classified.

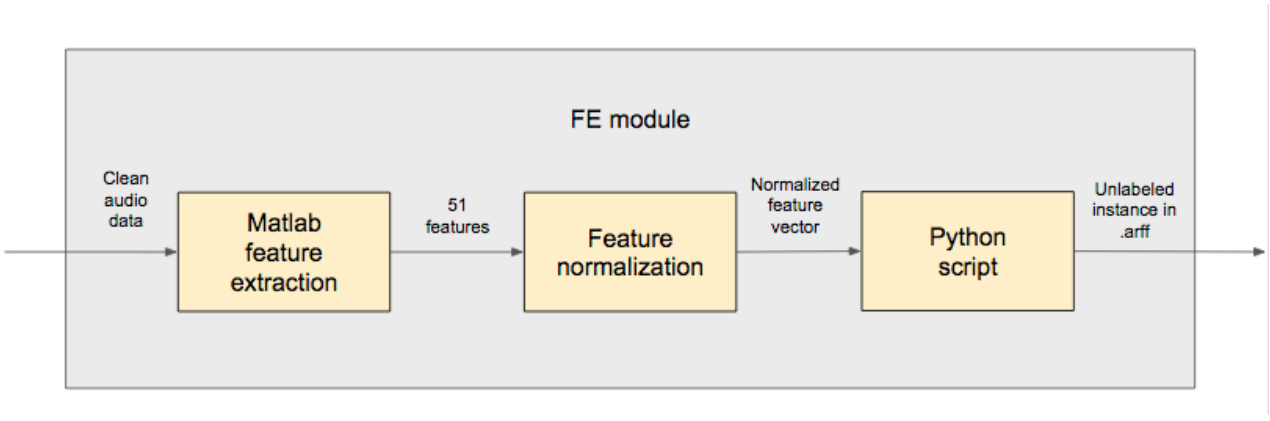

Figure 15: Feature extraction module

6.1.3 Affect Classification

The AC module receives the unlabeled instance and uses the MLP-based affect classification models generated in offline training, described in section 5.2, to makes a prediction about the corresponding valence and arousal. The classification is performed at a success rate of 79.4% and 84.4% for valence and arousal, based on the cross validation performance during offline training. A python script invokes the bash command provided by Weka to perform the classification on the unlabeled instance and parses the output to extract the predicted valence and arousal values and exports it into a log file along with the feature vector and a timestamp.

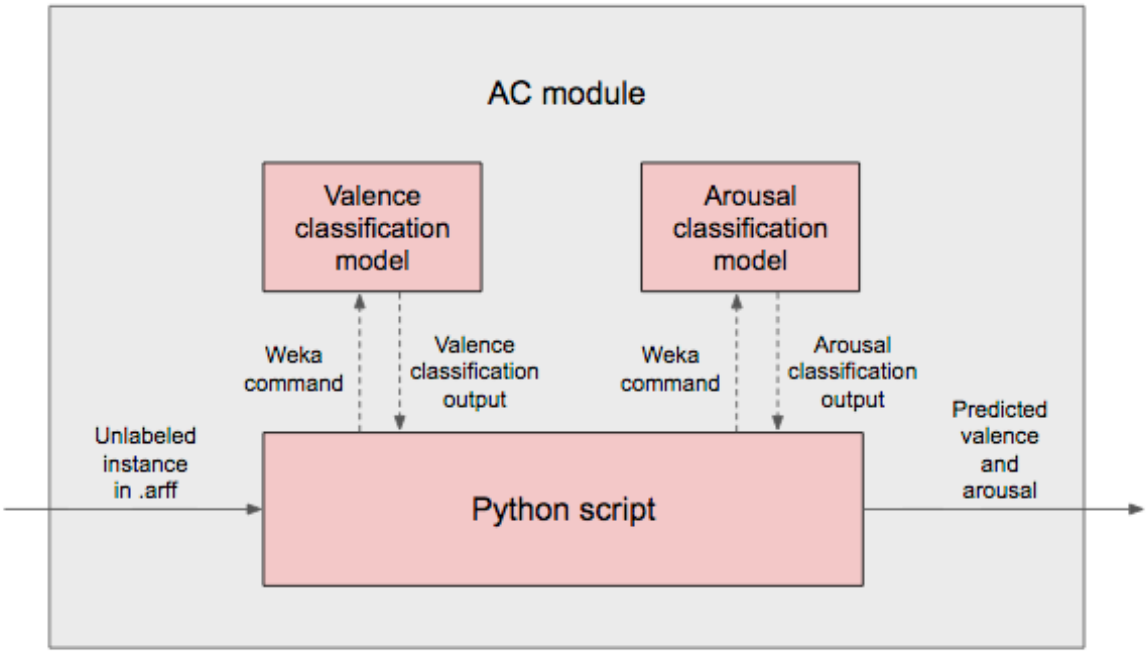

Figure 16: Affect classification module

6.2 Physical Setup

The physical setup of the interaction between NAO and a person is designed with consideration of the requirement of the body language modal and the emotion modeling components of the overall research. Even though these are not within the scope of this thesis, it is important that they are addressed because the vocal intonation modal is a subcomponent in the simulation of a natural HRI. As discussed in section 3.2.3, the NAO robot is placed on a table with a distance of

0.75 m from the speaker. The Voice tracker II Array Microphone is placed in front of NAO and the captured voice signal is sent to a desktop computer that is running the real-time vocal intonation affect recognition program, which is hidden from the view of the speaker. This Wizard of Oz (WOZ) operation is suitable for this application because the activity log of the affect recognition can be monitored during the interaction. In addition, during the multimodal interaction, the body language, multimodal affect recognition is also in operation which the processor on the NAO robot may not be able to handle. NAO can be communicated through Wi-Fi, which allows the emotion modeling to be performed on the desktop, which further decreases the work to be done on the robot itself. Figure 17 shows the physical setup during the real-time affect recognition of the vocal intonation modal.

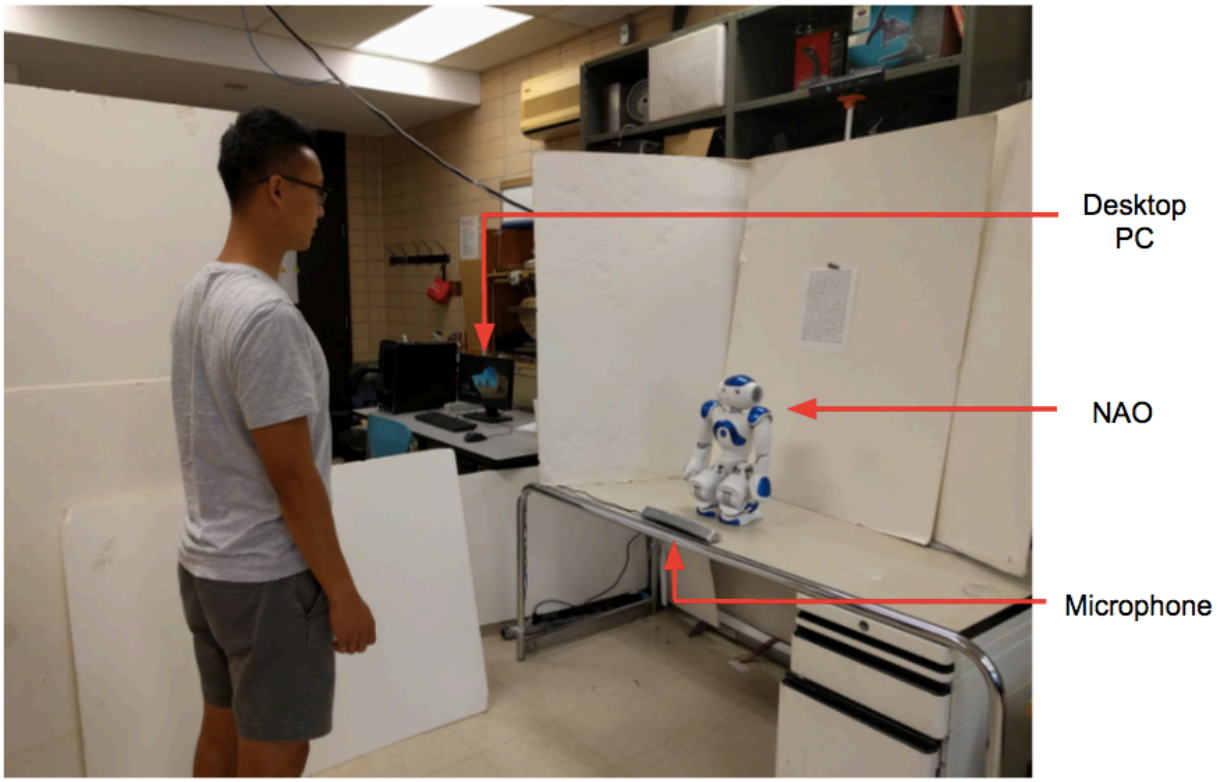

Figure 17: Physical setup of the real time vocal intonation affect recognition

Chapter 7
Discussion and Conclusion

7 Discussion and Conclusion

In this thesis, the process of the development of a system for real time affect recognition from vocal intonation modal was presented.

7.1 Discussion

7.1.1 Summary of Contributions

In order to built the affect recognition system, a database of emotional speech was created from three subjects. Feature extraction was performed in two ways to build the training data for the offline training of the valence and arousal classification models. The first was by using the QA5 SDK, which its resultant models produced cross validation results of 54.2% and 59.0% for valence and arousal, respectively. The second was by using the Matlab Audio Analysis Library, which its resultant models produced cross validation results of 79.4% and 84.4% for valence and arousal, respectively. The higher performance classification models from the second approach was used to classify the speaker speech for the real-time vocal intonation affect recognition.

7.1.2 Limitations

The classification rate of the final classification model is comparable or even better than human performance in affect recognition and therefore its integration to the multimodal system will introduce novelty to the overall affect recognition system. However, it is important to understand the limitations that comes along this result. The first limitation comes from the use of an external microphone for voice capturing. The interaction would be more natural if the NAO can hear the speaker's voice through its own set of microphones, an imitation of human ears. This can be justified by the primary objective of the research, which is to investigate the performance of affect recognition, because the ability of the microphones on NAO to capture the speaker voice is inferior to that of the Voice Tracker II Array Microphone. Nevertheless, the resultant model was trained to operate with an external microphone, which would not be available in an interaction between just human and NAO. The second limitation comes from the subject selection during

the emotional speech collection. The three subjects are all male in their early 20's, which is not very representative of the general population, in particular the elderly age group. This was due to the challenges in acquiring subject from outside of the ASB Lab, for example finding people who are willing to take part and preparing the appropriate ethics approval. The task of affect classification will become more complicated as the subject demographic becomes more reflective of that of the general public, thus, we cannot expect the same affect classification rate during real-life situations. Introducing more variance in the age and gender for the subject selection will lead to higher speaker independency.

7.1.3 Future Work

Despite the limitations of the final affect classification model, their identification sheds light on the direction of future work. For more natural HRI in real time, the voice capturing should be done with the microphones on the robot rather than using an external microphone, although it may hinder the quality of sound. In order to improve the classification rate of the affect state through vocal intonation, the subject selection must be done with a consideration for representing the specific target group. Additionally, feature selection and extraction can be experimented to further optimize the classification. In spite of these possible improvements, the final real time vocal intonation affect recognition system performs well for the purpose of this research and can be integrated with the body language modal to produce a multimodal affect recognition system. The future research beyond the scope of affect recognition can be the detection of wider range of affect states. Nine affect states were detected in this thesis but a human is capable of detecting more and this ability would allow robots to handle greater variety of social situation and build a closer relationship with people. Another direction of research could be the methodology in the deployment of the social robots in the real-world. The one of the main purposes of higher HRI quality is the acceptance of social robots in the live of people and its possibility at this stage in research should be investigated. Consequently, this initiative will further clarify the areas in HRI that requires more effort for people's fulfilling experience with social robots lead to its long-term immersion into the real world.

7.2 Conclusion

With high accuracy affect recognition, social robots will be able to engage in an effective and close interaction with people. The field of social assistive robots is still quite new and there are

still many areas that require improvement for natural HRI. Nevertheless, the interpretation of the human emotion is a fundamental ability that we as human possess and a significant step for robots to be able to coexist with human.

References or Bibliography

8 Works Cited

[1] D. Feil-Seifer and M. Mataric, "Socially Assistive Robotics," *IEEE Robotics & Automation Magazine,* vol. 18, no. 1, pp. 24-31, 2011.

[2] C. Breazeal, "Social Interactions in HRI: The Robot View," *IEEE Trans. Syst. , Man, Cybern. C,* vol. 34, no. 2, pp. 181-186, 2004.

[3] J. Terao, L. Trejos, Z. Zhang and G. Nejat, "An intelligent socially assistive robot for health care," in *IMECE*, Boston, 2008.

[4] M. P. Lawton, K. V. Haitsma and J. Klapper, "Observed affect in nursing home residents with Alzheimer's disease," *J. Gerontol. B Psychol. Sci. Soc. Sci.,* vol. 51, no. 1, pp. 69-78, 1996.

[5] K. Scherer, "Expression of emotion in voice and music," *Journal of Voice,* vol. 9, no. 3, pp. 235-248, 1995.

[6] L. Bosch, "Emotions: what is possible in the ASR framework," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[7] R. Picard, Affective computing, vol. 252, Cambridge: MIT press, 1997.

[8] C. Breazeal, Designing sociable robots, MIT press, 2004.

[9] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies,* vol. 59, no. 1-2, p. 157–183, 2003.

[10] B. Schuller, S. Reiter and G. Rigoll, "Evolutionary feature generation in speech emotion recognition," in *IEEE International Conference on Multimedia and Expo ICME 2006*, 2006.

[1] K. Scherer, "Vocal Affect Expression: A Review and a Model for Future Research," *Psychological Bulletin,* vol. 99, no. 2, pp. 143-165, 1986.

[2] C. Darwin, The expression of the emotions in man and animals, Chicago: The University of Chicago Press, 1965.

[3] K. Scherer and J. Oshinsky, "Cue utilization in emotion attribution from auditory stimuli," *Motivation and Emotion,* vol. 1, no. 4, p. 331–346, 1977.

[4] P. Silvia, Exploring the Psychology of Interest, New York: Oxford University Press, 2006.

[5] K. Scherer, Facets of Emotion: Recent Research, New York: Psychology Press, 1988.

[6] Y. Matsuda, T. Fujimura, K. Katahira, M. Okada, K. Ueno, K. Cheng and K. Okanoya, "The implicit processing of categorical and dimensional strategies: an fMRI study of facial emotion perception," vol. 7, no. 1, p. 551, 2013.

[7] T. Eerola and J. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music,* vol. 39, no. 1, pp. 18-49, 2010.

[8] K. Scherer, "Psychological Models of Emotion," *The Neuropsychology of Emotion,* pp. 137-162, 2000.

[9] G. Paltoglou and M. Thelwall, "Seeing Stars of Valence and Arousal in Blog Posts," *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING,* vol. 4, no. 1, pp. 116-123, 2013.

[10] Technical University of Berlin, Institute of Speech and Communication, department of communication science, "Berlin Database of Emotional Speech," [Online]. Available: http://emodb.bilderbar.info/docu/.

[11] A. P. M. R. W. S. B. W. F. Burkhardt, "A Database of German Emotional Speech".

[2] Ryerson Multimedia Research Laboratory, "RML Emotion Database | RML | Ryerson Multimedia Research Laboratory," [Online]. Available: http://www.rml.ryerson.ca/rml-emotion- database.html. [Accessed 3 October 2015].

[2] D. V. a. C. Kotropoulos, "A State of the Art Review on Emotional Speech Databases," Artificial Intelligence & Information Analysis Laboratory.

[2] S. Koelstra, "SEMAINE Database - Home," 2015. [Online]. Available: http://semaine-db.eu/. [Accessed 9 Oct 2015].

[2] I. N. K. D. Terrence Fong, "A survey of socially interactive robots," *Robotics and Autonomous Systems,* vol. 42, p. 143–166, 2003.

[2] H. Siegel, "Opensource 3D printed “Poppy” humanoid enables experimentation in robot design," 2015. [Online]. Available: http://robohub.org/3d-printed-opensource-poppy-robot-hopes-to-further-experimentation-in-morphology/. [Accessed 23 Aug 2016].

[2] SoftBank Robotics, "https://android.aldebaran.com/doc/index.html," Aldebaran, 2016. [Online]. [Accessed 24 Aug 2016].

[2] Aldebaran, "Who is NAO?," SoftBank Robotics, 2016. [Online]. Available: https://www.ald.softbankrobotics.com/en/cool-robots/nao. [Accessed 24 August 2016].

[2] Nemesysco Ltd., "QA5 SDK Product Description & User Guide," Netanya.

[3] A. Eriksson and F. Lacerda, "Charlatanry in forensic speech science: A problem to be taken seriously," *The International Journal of Speech, Language and the Law,* vol. 14, no. 2, pp. 169-193, 2007.

[3] A. Liberman, "Apparatus and Methods for Detecting Emotions,". US Patent 6,638,217 B1, 28 Oct 2003.

[32] G. T., "Matlab Audio Analysis Library," Matlab, 18 March 2014. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/45831-matlab-audio-analysis-library. [Accessed 2016].

[33] T. L. Nwe, S. W. Foo and L. C. D. Silva, "Detection of stress and emotion in speech using traditional and FFT based log energy features," *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on,* vol. 3, pp. 1619-1623, 2003.

[34] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *The Journal of the Acoustical Society of America,* vol. 52, no. 4, pp. 1238-1250, 1972.

[35] Machine Learning Group at the University of Waikato, "Weka 3: Data Mining Software in Java," [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/.

[36] S. G. S. A. M. a. E. S. Brandon Heenan, "Interaction, Designing Social Greetings in Human Robot," Calgary.

[37] S. W. Smith, "Moving Average Filters," in *The Scientist & Engineer's Guide to Digital Signal Processing*, California Technical Pub, 1998.

[38] T. Vogt and E. André, "An Evaluation of Emotion Units and Feature Types for Real-Time Speech Emotion Recognition," *Künstliche Intelligenz,* vol. 25, no. 3, pp. 213-223, 2011.

[39] J. Chan, "READ ME," Toronto, 2011.

[40] D. Belsley, "Collinearity and weak data in regression," in *Conditioning diagnostics*, New York, Wiley, 1991.

[41] Y. Cheng and L. Che-hui, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Systems with Applications,* vol. 36, no. 2, p. 2473–2480, 2009.

[42] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[43] E. Frank, Y. Wang, S. Inglis, G. Holmes and I. Witten, "Using Model Trees for Classication".

[44] J. Bancroft, "Separation of voiced and unvoiced speech, and silence, using energy and periodicity," *The Journal of the Acoustical Society of America,* vol. 68, no. 1, p. S70, 1980.

[45] K. S. A. B. B. B. Bachu R.G., "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal".

[46] P. J. and A. Přibilová, "Statistical Analysis of Complementary Spectral Features of Emotional Speech in Czech and Slovak," *Springer-Verlag Berlin Heidelberg,* p. 7, 2011.

[47] A. KOVAC, M. M. HALAS, P. P. PARTILA and M. VOZNAK, "Impact of Emotions on Fundamental Speech Signal Frequency," *Latest Trends in Information Technology,* 2012.

[48] G. Peeters, "A Large Set of Audio Features for Sound Description," Paris, 2004.

[49] J. P. Bello, "Low-level features and timbre," New York.

[50] B. Schuller, G. G. Rigoll and J. Stadermann, "Affect-Robust Speech Recognition by Dynamic Emotional Adaptation," *Speech Prosody,* 2006.

[51] D. Giannoulis, M. Massberg and J. D. Reiss, "Automating Dynamic Range Compression," *Journal of the Audio Engineering Society,* vol. 61, no. 10, October 2013.

[52] J. H. L. H. Seyed Omid Sadjadi, "Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux," *IEEE SIGNAL PROCESSING LETTERS,* vol. 20, no. 3, pp. 197-200, March 2013.

[53] S. C. D. J. P. Chandrasekar, "Emotion Recognition from Speech using Discriminative Features," *International Journal of Computer Applications,* vol. 101, no. 16, pp. 31-36, 2014.

[54] M. B. a. I. E. Lindasalwa Muda, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," *Journal of Computing,* vol. 2, no. 3, pp. 138-143, March 2010.

[55] R. S. a. H. N. Andras Zolnay, "ACOUSTIC FEATURE COMBINATION FOR ROBUST SPEECH RECOGNITION," Aachen.

[56] ccrma.stanford.edu, "Hamming Window," [Online]. Available: https://ccrma.stanford.edu/~jos/sasp/Hamming_Window.html. [Accessed 23 Aug 2016].

[57] L. D. Enochson and R. K. Otnes, "Programming and Analysis for Digital Time Series Data," 1968.

[58] MathWorks, "hamming," Matworks, [Online]. Available: http://www.mathworks.com/help/signal/ref/hamming.html. [Accessed 23 Aug 2016].

[59] "5.4 Filterbank Analysis," [Online]. Available: http://izanami.tl.fukuoka-u.ac.jp/SLPL/HMM/HTKBook/node59.html. [Accessed 24 August 2016].

[60] CMUSphinx, "Frequenty Asked Questions (FAQ)," 25 May 2016. [Online]. Available: http://cmusphinx.sourceforge.net/wiki/faq#qwhat_speech_feature_type_does_cmusphinx_use_and_what_do_they_represent. [Accessed 2016].

[61] A. Robel, "Fundamental frequency estimation," Berlin, 2006.

[6 2] S. F. L. D. S. T. Nwe, "Speech emotion recognition using hidden Markov models," *Speech Communication,* vol. 41, no. 4, pp. 603-623, 2003.

[6 3] A. S. P. S. A. Wankhammer, "CHROMA AND MFCC BASED PATTERN RECOGNITION IN AUDIO FILES UTILIZING HIDDEN MARKOV MODELS AND DYNAMIC PROGRAMMING," *Int. Conference on Digital Audio Effects,* 2009.

[6 4] H. K. J. Lee, "On the Importance of Tonal Features for Speech Emotion Recognition," *Journal of Broadcast Engineering,* vol. 18, no. 5, pp. 713-721, 2013.

[6 5] J. P. Bello, "Chroma and tonality".

[6 6] P. R. a. A. Drygajlo, "Entropy Based Voice Activity Detection in Very Noisy Conditions," *threshold,* vol. 5, no. 5.5, 2001.

[6 7] D. H. a. S. Krishnan, "On the Use of Complementary Spectral Features for Speaker Recognition," *EURASIP Journal on Advances in Signal Processing,* no. 1, 2008.

[6 8] P. H. Martin Rocamora, "Comparing audio descriptors for singing voice detection in music audio files," Montevideo, 2007.

[6 9] S. C. a. D. J. P. Chandrasekar, "Emotion Recognition from Speech using Discriminative Features," *International Journal of Computer Applications,* vol. 101, no. 16, pp. 31-36, 2014.

[7 0] L. Rabiner, "Digital Speech Processing - Linear Predictive Coding (LPC)".

[7 1] N. Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition," *International Journal For Advance Research in Engineering And Technology,* vol. 1, 2013.

[7 2] C. S. Sergey Ioffe, "Batch Normalization: Accelerating Deep Network Training b y Reducing Internal Covariate Shift," arXiv preprint, 2015.

[7 A. K. R. C. D. L. S. J. W. Wenyi Zhao, "Discriminant analysis of principal components for face recognition," *Face Recognition,* vol. 163, pp. 73-85, 1998.

[7 "Neural network models (supervised)," [Online]. Available: http://scikit-learn.org/dev/modules/neural_networks_supervised.html.

[7 Y. Chu, "A neural network which learns decision boundaries with nonlinear clustering," in *1991 IEEE International Joint Conference on Neural Networks*, 1991.

[7 "Comparing supervised learning algorithms," 2015. [Online]. Available: http://www.dataschool.io/comparing-supervised-learning-algorithms/. [Accessed 8 June 2016].

[7 S. Koelstra, "SEMAINE Database - Home," 2015. [Online]. Available: http://semaine-db.eu/. [Accessed 9 Oct 2015].

[7 E. F. L. T. R. K. Gordon Paynter, "Attribute-Relation File Format (ARFF)," 1 November 2008. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/arff.html. [Accessed 23 Aug 2016].

Appendices

Excerpt from Snow White

The following excerpt from the short story Snow White [77] was read for the creation of the database of emotional speech in Chapter 3.

> Once upon a time there lived a lovely princess with fair skin and blue eyes. She was so fair that she was named Snow White. Her mother died when Snow White was a baby and her father married again. This queen was very pretty but she was also very cruel. The wicked stepmother wanted to be the most beautiful lady in the kingdom and she would often ask her magic mirror, “Mirror! Mirror on the wall! Who is the fairest of them all?” And the magic mirror would say, “You are, Your Majesty!” But one day, the mirror replied, “Snow White is the fairest of them all!” The wicked queen was very angry and jealous of Snow White. She ordered her huntsman to take Snow White to the forest and kill her. “I want you to bring back her heart,” she ordered. But when the huntsman reached the forest with Snow White, he took pity on her and set her free. He killed a deer and took its heart to the wicked queen and told her that he had killed Snow White. Snow White wandered in the forest all night, crying.