CLASSIFICATION OF STRESS LEVEL BASED ON SPEECH FEATURES

ARSHED AHMED JASIM

UNIVERSITI UTARA MALAYSIA

2014

CLASSIFICATION OF STRESS LEVEL BASED ON SPEECH FEATURES

A dissertation submitted to Dean of Research and Postgraduate Studies Office

In partial Fulfilment of the requirement for the degree

Master of Science (Information Technology)

Universiti Utara Malaysia

By

Arshed Ahmed Jasim

Permission to Use

In presenting this dissertation in fulfilment of the requirements for a Master of Science in Information Technology (MSc. IT) from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this dissertation in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my dissertation.

Requests for permission to copy or to make other use of materials in this dissertation, in whole or in part should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

Abstrak

Kehidupan kontemporari adalah penuh dengan cabaran, gangguan, tarikh akhir, kekecewaan dan permintaan yang tidak berkesudahan. Ini boleh mengakibatkan sesorang itu stres. Stres telah menjadi satu fenomena global yang dialami dalam kehidupan moden harian kita. Stres mungkin memainkan peranan penting dalam gangguan psikologi dan tingkah laku seperti kebimbangan atau kemurungan. Oleh itu, pengesanan awal tanda-tanda dan gejala stres merupakan penawar ke arah mengurangkan kesan buruk dan kos yang tinggi dalam pengurusan stres. Usaha penyelidikan yang dibentangkan ini merangkumi teknik Pengenalan Percakapan Automatik (ASR) untuk mengesan stres sebagai alternatif yang lebih baik berbanding pendekatan yang lain seperti analisis kimia, kekonduksian kulit, elektrokardiogram yang mahal dan mempunyai kesan halangan dan gangguan. Dua set data suara direkodkan daripada sepuluh orang pelajar Arab di Universiti Utara Malaysia (UUM) iaitu dalam mod rehat dan stress. Ciri-ciri percakapan seperti frekuensi asas (f_0); formants (F1, F2, dan F3), tenaga dan Pekali Frekuensi Cepstral Mel (MFCC) ini diekstrak dan dikelaskan menggunakan jiran K-terdekat, Analisisa Diskriminan Linear dan Rangkaian Neural Buatan. Keputusan dari nilai purata frekuensi asas mendedahkan bahawa peningkatan stres adalah berkait rapat dengan pertambahan nilai frekuensi asas. Daripada tiga metod pengkelasifan, prestasi jiran K-terdekat (KNN) adalah terbaik diikuti oleh analisisa diskriminan linear (LDA) manakala rangkaian neural buatan (ANN) menunjukkan prestasi yg paling rendah. Klasifikasi tahap stres rendah, sederhana dan tinggi telah dilakukan berdasarkan keputusan klasifikasi daripada KNN. Kajian ini menunjukkan kebolehgunaan maju ASR sebagai cara yang lebih baik pengesanan stres dan pengkelasan.

Abstract

Contemporary life is filled with challenges, hassles, deadlines, disappointments, and endless demands. The consequent of which might be stress. Stress has become a global phenomenon that is been experienced in our modern daily lives. Stress might play a significant role in psychological and/or behavioural disorders like anxiety or depression. Hence early detection of the signs and symptoms of stress is an antidote towards reducing its harmful effects and high cost of stress management efforts. This research work thereby presented Automatic Speech Recognition (ASR) technique to stress detection as a better alternative to other approaches such as chemical analysis, skin conductance, electrocardiograms that are obtrusive, intrusive, and also costly. Two set of voice data was recorded from ten Arabs students at Universiti Utara Malaysia (UUM) in neural and stressed mode. Speech features of fundamental, frequency (f_0); formants (F1, F2, and F3), energy and Mel-Frequency Cepstral Coefficients (MFCC) were extracted and classified by K-nearest neighbour, Linear Discriminant Analysis and Artificial Neural Network. Result from average value of fundamental frequency reveals that stress is highly correlated with increase in fundamental frequency value. Of the three classifiers, *K*-nearest neighbor (KNN) performance is best followed by linear discriminant analysis (LDA) while artificial neural network (ANN) shows the least performance. Stress level classification into low, medium and high was done based of the classification result of KNN. This research shows the viability of ASR as better means of stress detection and classification.

Acknowledgement

By the Name of Allah, the Most Gracious and the Most Merciful

First and foremost, I thank to Allah for blessing me with good health to be able to complete this project. This dissertation is accomplished by the student but the knowledge and experiences gathered, evolved during the process through the efforts of many people. Without their cooperation, encouragement and suggestion; this study would not have been possible.

I heartily thank my supervisors Dr. Shahrul Azmi Mohd.Yusof and Ms. Aniza Mohamed Din They have supported me throughout my research process. They gave support, insight, guidance and encouragement throughout to fulfil my study.

My acknowledgements would not be complete until I thank my great father and my lovely mother for their prayers, love, and support in the duration of my study in Malaysia. Who sacrificed much and supported my efforts with understanding and constant encouragement. Without them, it is almost impossible for me to complete this master's degree. May Allah bless them.

I would like to express my gratitude to my brother and my sister for their love and moral care. They are supporting me spiritually throughout my life. Without them, I'm nothing here.

Finally, many thanks go to all of Malaysian people, especially UUM lecturers and staffs, for their very good dealing with all of international students. As well as, Thanks to College of Arts and Science with its community whom has made this possible by organizing this course which provided the opportunity for me to share and learn their information in applying the essential apparatus to the mission.

"Thank you UUM"

Table of Contents

List of Tables

Table **Page**

List of Figures

Figure **Page**

List of Acronyms

A	Age of the speakers
ANN	Artificial Neural Network
ANS	Automatic Nervous System
ASR	Automatic Speech Recognition
BN	Bayesian Network
BPNN	Back-propagation Neural Network
BVP	Blood Volume Pressure
CB	Critical Band
CRs	Classification Rates
CV	Consonant-Vowel
DFT	Discrete Fourier Transform
DTW	Discrete Wavelet Transform
f0	Fundamental Frequency
F1, F2, F3	First, Second and Third Formants
F, M	Female, Male
FFT	Fast Fourier Transform
FM	Frequency Modulation
FS, FN	Female Stress, Female Neutral
GMM	Gaussian Mixture Model
GSR	Galvanic Skin Response
HCI	Human- Computer Interaction
HCNN	Hidden Control Neural Network
HMM	Hidden Markov Model
HR	Current Heart Rate

KNN	K-Nearest Neighbor
l	Exertion level
LDA	Linear Discriminant Analysis
LM	Levenberg-Marquardt back propagation
LPC	Linear Predictive Coding
MAP	Maximum A Posteriori
MFCC	Mel-Frequency Cepstral Coefficient
MHR	Maximum Heart Rate
MLP	Multi-Layer Perceptron
MS, MN	Male Stress, Male Neutral
NN	Neural Network
PAD	Pitch, Amplitude, Duration
PLP	Perceptual Linear Prediction
RHR	Normal Heart Rate
RNN	Recurrent Neural Network
ROS	Rate of Speech
SLM	Sound Level Meter
ST	Skin Temperature
STT	Speech to Text
SUSAS	Speech under Simulated and Actual Stress
SVM	Support Vector Machine
TEO	Teager Energy Operator
VQ	Vector Quantization
WER	World Error Rate

CHAPTER ONE

INTRODUCTION

1.0 Introduction

This section serves as a broad introduction to the study. It contains the background of the study, motivation, and problem statement. In addition, it also presents research questions and the objectives of the research, the scope and significance of the study.

1.1 Background

Contemporary life is filled with challenges, hassles, deadlines, frustrations, disappointments, and endless demands. The consequent of which might be stress. Stress has become a global phenomenon that is been experienced in our modern daily lives (Lu et al., 2012). For many people, stress is so commonplace – in traffic, markets, schools, or at work that it has become a way of life so much that ability to cope with stress is seen as a plus quality. While to some stress is a nightmare. Stress is not always bad, in small doses, it can help propel and motivate an individual under pressure to do better (Dhole & Gurjar, 2013). But being constantly running in emergency mode (stressed), the body and mind might pay the price. Affirming this is the studies report that stress might play a significant role in psychological and/or behavioural disorders like anxiety or depression (Dhole & Gurjar, 2013; Lu et al., 2012). Early detection of the signs and symptoms of stress is an antidote towards reducing its harmful effects and high cost of stress management efforts. Ability to detect stress and the level can be of use vital in applications that are stress sensitive such as

psychological testing voice activated military equipment, and deception detection (Dhole & Gurjar, 2013).

Though there are several means of stress detection such as chemical analysis, skin conductance, electrocardiograms, etc.(Lu et al., 2012). These methods are however expensive, cumbersome, ineffective, and also intrusive. There is therefore the need for effective, convenience, easy to use and cost effective means of stress detection. According to studies, speech contains enormous information (Narayana & Kopparapu, 2009) and that clean acoustic signal contain about 25% information about the speaker (Dhole & Gurjar, 2013). This has made speech a viable means of determining emotion or stress level of humans. The most convenient way of detecting stress is through speech analysis with automatic speech recognition. Speaker's characteristics such as personality, emotion, response to situations such as fatigue, stress, and medical conditions constitute major part of speech signal (Sigmund, 2010).

Stress has enormous negative effect on individual and public health concerns, hence the need to device a stress detection approach that is not only automatic but also ubiquitous (Dhole & Gurjar, 2013). Automatic and ubiquitous stress detection will facilitate easy and early stress detection, and hence it's early management. Likewise, it will enable health workers the ability to observe the degree and spread of stress within the populaces. Of equal interest to the ASR researchers is the performance of automatic speech recognition (ASR) that degrades considerably when the condition in which it was trained differs from the testing condition (Amuda, Boril, Sangwan, & Hansen, 2010; Zhou, Hansen, & Kaiser, 2001). One of such conditions is stress. This fact is acknowledged by researchers such as (Bou-Ghazale & Hansen, 2000). In ASR research and particularly in this work, the definition of stress as given

by Murray, Baber, and South (1996) “Stress is an effect on the production of speech (manifested along a range of dimensions), caused by exposure to a stressor” will be the working definition of stress of this dissertation. Stress is a psychological state that is a response to a perceived threat or task demand and is normally accompanied by specific emotions. As stated by Hong, Ramos, and Dey (2012), the presence of a stressor implies that the speech is stressed will be the assumption of this dissertation.

Stress has become a silent killer disease affecting substantial number of people globally (Costello et al., 2009). It is a cause of wide range of diseases such as immune deficiencies, cerebrovascular disease, diabetes, cardiovascular disease, and a source of sudden deaths (Kurniawan, Maslov, & Pechenizkiy, 2013). Of recent, well-being and health-care researchers has focused on stress management going by a broader recognition of potential problems caused by chronic stress (Bakker, Holenderski, Kocielnik, Pechenizkiy, & Sidorova, 2012). The most effective means of stress management is its early detection. Hence the need to device a stress detection approach that is not only automatic but also ubiquitous (Bakker et al., 2012; Dhole & Gurjar, 2013; Lu et al., 2012). Though there are several approaches to stress detection and measurement such as physiological approach: blood volume pressure (BVP), galvanic skin response (GSR), and skin temperature (ST) (Dhole & Gurjar, 2013; Zhai & Barreto, 2008). Others include facial affect, body gesture and visual aesthetic (Lu et al., 2012). All the above approaches to stress detection are however, expensive, cumbersome, intrusive, and non-convenience and above all their use can also induce stress (Dhole & Gurjar, 2013; Sigmund, 2010). Hence the need for stress detection approach that is cheaper, convenient, and non-intrusive, ubiquitous, and above all effective. Fulfilling these requirement is speech signal approach (Dhole & Gurjar, 2013). Non-intrusive

technologies that automatically recognize stress can serve a potent tool to detect and monitor stress levels(Bakker et al., 2012; Kurniawan et al., 2013).

Most of the available stress speech corpus such as SUSAS database contains speech under simulated and acted stress which makes the corpus unreliable. This absence of a general and reliable stressed speech corpus (Wang, 2009) has motivated researchers to collect a more constrained but more reliable stressed speech corpus.

The application of Machine Learning approach to recognition of stress states is a relatively new technique that has found wider usage. Software used include Weka and Matlab (Casale, Russo, Scebba, & Serrano, 2008). Although there are several varieties of linear and non-linear classifiers used by researchers for speech vowel recognition, the most widely used are Neural Network based classifiers of Multi-Layer Perceptron, Levenberg-Marquardt (LM) and Support vector machine (SVM), Linear Discriminant Analysis (LDA) and K-Nearest Neighbor (KNN). For the purpose of this dissertation, three non-linear classifiers of Levenberg Marquart trained Neural Network (LM), K-Nearest Neighbors (KNN), and a linear classifier of Linear Discriminant Analysis (LDA) will be used for classifying stress level in speech.

1.2 Problem Statement

Although ASR has recorded a greater progress in few domains, its performance is still less than expected (Wang, 2009). Performance of present ASR degrades considerably when exposed to noise, distortion and stressed speech (Sigmund & Dostal, 2004; Wang, 2009). Speech variability due to stress is far greater on speech production than variability across

speakers (Benzeghiba *et al.*, 2007). Physical task stress has been shown to significantly impact recognition performance (Chen, 1988; Hansen, 1988, 1989; Paul, 1987; Rajasekaran et al., 1986). The fact that ASR has not recorded wide usages was first predicated due not only to limitations of the currently available technology, but also to human factor issues that are peculiar to usage of ASR systems (Entwistle & Adviser-Granaas, 2005).

Some of the studies have used prosodic information of the speech at different class-level (sentences, word or syllable) (Koolagudi et al.2011; Zhang, 2012; Rao et al., 2013). The most commonly used spectral features are Mel-Frequency Cepstral Coefficients (MFCC) (Han et al., 2012; Koolagudi, Kumar & Rao, 2011; Khanna & Kumar, 2011; Mao et al., 2009). Patil and Hanseen (2008) have used MFCC at frame level while they achieved 62% to 66% accuracy. Dhole et al. (2013) have used speech duration and amplitude features for classification of stress and 84% accuracy is reported but Dhole et al. (2013) used only low frequency components at frame level. Lu et al. (2012) have used 20-MFCC features with energy, pitch and speaking rate reported 76% results at frame level. Moreover according to Lu et al. (2012) MFCC is generic speech feature that is widely used in speech analysis. According to He et al., (2009) stress classification results using speech at vowel level have very similar accuracy for GMM and KNN that is 55.34% to 73.76%. Very few studies have used MFCC features with KNN and LDA for stress detection (chee et al., 2009; Torabi et al., 2008). Furthermore, very few studies classified different levels of stress using speech processing. All of the studies listed above are done using English speaking native speakers. None have used speakers from Arab countries which may have significantly different English accent. This study uses LDA and KNN for stress levels classification at word level while using MFCC features to analyse the efficiency of different stress level using speech samples obtained from speakers of Arab origins.

1.3 Research Questions

In the cause of this study, answers will be provided to the following research questions:

i. How does physical stress affect speech production?
ii. What speech feature is most affected by physical stress?
iii. How can the stressed levels be effectively classified?

1.4 Research Objectives

The main objective of this dissertation is to probe how stress affects speech characteristics with specific aim of detecting and classifying stressed levels. This will be achieved by the following sub objectives:

(i) To analyse stressed speech characteristics due to physical task so as to identify stress variant features of speech using Mel Frequency Cepstral Coefficients (MFCC) and other acoustic features extracted from collected speech corpus.

(ii) To identify speech feature most affected by physical induced stress.

(iii) To classify physical induced stress into different categories of Low (L), Medium (M), and High (H) and comparing the performance of classifiers such as K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA) and Levenberg-Marquardt Neural Network (LM).

1.5 Motivation

Stress has become a daily life occurrence in our modern life. Stress impacted negatively on long term physical and mental wellness of individual (Lu et al., 2012). Remarked by Bakker

et al., (2012) that stress management should start in earnest before it escalates to illness causing levels motivates this work. A necessary prerequisite for early detection of stress is the development of cheap, easy to use, ubiquitous, and widely available method of stress detection. Of all method of stress detection, speech is the most suitable method. Furthermore, ability to characterized and model speech under stress condition is a means towards attainment of stress adaptive ASR.Therefore, the necessity for exploring the characteristics of acoustical features of the speech under stress and formulating the robust compensation techniques activates the motivation for building a more powerful speaker independent recognition system that is insensitive to emotional mood and stress. The formulation of a more powerful tool for robust speech recognition needed to reach the goal of better performance and improved recognition capability of speech produced under stressful environment.

1.6 Contributions

The major contribution of this dissertation is the identification of stress variant features of stressed speech. Other contributions include:

i. Collection of speech corpus under neutral and stressed condition of physical exercise.
ii. Identification of the most suitable feature extraction algorithm for stressed speech.
iii. Evaluation of different classifier for based on classification rates (CRs) of stressed speech.

1.7 Scope

The scope of this dissertation shall be on the intrinsic and short-term variability in speech production resulting from physical workload induced stress for identification and subsequent classification of stress levels based on the speech corpus collected from Arab students in the Universiti Utara Malaysia (UUM) due to time constraint and readiness to partake in the experiment.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

In this section, related literatures to this study are reviewed. The impact of variability on ASR performance, types and sources of variability are discussed. Stress, types of stress, and stressors were discussed also.

2.1 ASR performance and variability

ASR performance degrades significantly due to variability as a result of difference in acoustic model (training condition) and the test data (Benzeghiba et al., 2007). This has impacted greatly on the performance of ASR and its wider application in real life environments (Hansen, Sangwan, & Kim, 2012). However, in response to the environmental condition (e.g. Noise/Lombard effect), and or speaker state of mind (e.g. emotion, stress, and health), humans modified their mode of speaking to reflect these situations (Hansen et al., 2012). The fact that in neutral/normal situations, human speech generation is rarely the same, further aggravated speech variability. Furthermore ASR are continuously being deployed in environments that are considerably different from which they were trained and hence taking into cognizance different forms of variabilities such as those of acoustic environment encountered in daily lives and those inherent in individual speakers such as stress when designing ASR is a prelude towards attainment of high performance ASR (Hansen et al., 2012; Lu et al., 2012).

2.2 Types of variability

The fundamental problem of speech recognition like any other pattern recognition problem is variability, which results in low recognition rate (higher WER). Sources of speech variability include duration, spectral, speaker, accent, stress, emotion, contextual, and noise. However, the most challenging of this variability includes accent, stress, and background noise (Yan, Vaseghi, Rentzos, & Ho, 2007). Sources of speech variability can be broadly categorized into two: speaker's intrinsic characteristics and environmental sources (You & Adviser-Alwan, 2009). Longitudinally, speech variability can be further divided into long and short term(Godin, Hansen, Busso, & Katz, 2009).

2.2.1 Intrinsic variability in speech

Intrinsic variation in speech is due to factors that are directly related to speaker's characteristics. Such factors include: gender, age, rate of speech (ROS), accent, dialect, stress, and emotion (Benzeghiba et al., 2007; Yan et al., 2007). This research will focus on stress as a source of speech variability.

2.2.2 Environmental variability

Before now, the main source of variability is ASR is attributed to the environmental noise. This is evident in the number of publications related to environmental/noise robustness. Lately, distortion as a result of transmission channels and reverberation has been identified as an environmental factor of speech variability. Several methods are adopted in mitigating the environmental effect of ASR performance. Speech enhancement is aimed at generating a clean speech input signal devoid of environmental contaminations. Speech enhancement

approaches includes using noise-cancelling microphone or microphone array (Martin, 2005; Wu & Wang, 2006). Approaches in counteracting reverberation effects are proposed by (Nakatani, Juang, Kinoshita, & Miyoshi, 2005; Neely & Allen, 1979; Yegnanarayana & Murthy, 2000).

2.2.3 Long term variability

Long term variability in speech is closely related to speaker characteristics and hence speaker's dependent such as age, sex, accent or dialect. These speech characteristics are acquired and developed over a long period of term and remain stable after their full development (Godin et al., 2009).

2.2.4 Short term variability

Short term variability is due to factors outside the speaker's control. They are majorly induced by short term changes in the environment such as emotion, stress, illness, and environmental conditions (Benzeghiba et al., 2007). The focus of this work will be on short term induced variability.

2.3 Stress as a source of variability

Stress has a greater distortion effect on acoustic signal. Consequently, any circumstance that led a speaker to change speech production from neutral conditions becomes known as stress. (Wang, 2009). Based on their effects, stress can group into two classes of: perceptual and physiological. Variation or changes in the normal or neutral speaking environment result into perceptual stress. Sources of perceptual stress includes emotion, Lombard effect and task load. While physiological stress is as a result of physical impact on human body, such as

vibration, air density/pressure, drug interactions, and sickness (Narayana & Kopparapu, 2009). The model of changes induced by stress speech production as presented by Sigmud and Dostal (2004) depicted in Figure 2.1will form the basis of this research work. Based on figure 2.1, emotion stimulus such as physical exercise will lead into physiological changes (e.g. heart rate) that cause changes in the musculature (increased muscle tension). The increased in muscle tension affect the vocal track kinematics which consequently affect the quality of the speech produced by a speaker.

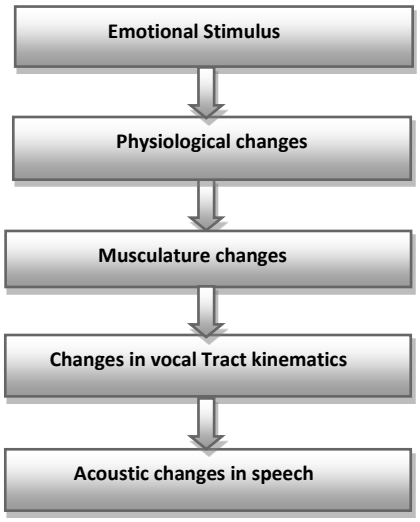


Figure 2.1. Modelling of changes caused by Emotion in Speech.

It is widely known that stress has a long term imprints on mental and physical well-being on individuals among which is changes in speech production (Lu et al., 2012). Among physiological effects of stress includes but not limited to: increase in the heart rate, respiration, and tension in the muscular tissues. The quality of speech production degraded considerably due to tension in the muscular tension on vocal cords and tract (Sigmund, 2010). Hence the viability of detecting stress through speech analysis. Stressed speech differs considerably from the neutral speech. This variability causes a remarkable performance degradation of ASR (Godin et al., 2009).

2.4 Types of stress and stressor

Stressed speech is brought about as a result in changes is speech production mechanism that differs from the neutral form. Stressed speech includes speech divergent in rate (fast, slow), effort (loud, soft), task demands (cognitive demands, physical task, fatigue), and under chemical inducement (drugs, alcohol) (Chen, (1988) and Hansen, (1988) in (Godin et al., 2009)).“Stress is observable variability in certain speech features due to a combination of unconscious response to stressors and/or conscious control” (Murray, Baber, & South, 1996). Stressors are factors that induced or caused stress. There is no clear distinction between stress types and stressor as stress are directly related to the stressor (Murray et al., 1996). Nonetheless stressor can be classified into chemical (drugs, alcohols, narcotics,), physical (exercise, noise, vibration), physiological (illness, workload, emotion, stress, anxiety, depression). In trying to clarify stress and stressor, (Cohen, Kessler, & Gordon, 1995) stated that stress can be describe as a subjective occurrence, while stressors are the observable or fictional actions and also stimuli that induced stress. The focus of this work shall be physical induced stress. Table 1 below adopted from (Sigmund, 2010) gives classification of stress and stressors.

As depicted in table 1, stressor may range from overcrowding, traffic congestion, violence, bereavement, redundancy, or unemployment to physical, chemical, biological, or psychological insults. Whether the person can adapt to or cope with the stress will depend on the nature and severity of the stressor and the person’s physical and mental state, which in turn depends on genetic, experiential, social, and environmental factors. From the table 1, stressor can be classified into an increasing order ranging from 0 - 3 based on its source and effects on human.

Table 2.1

Classification of Stressor in Speech Production System

Stress order	Classification	Stressors
0	**Physical**	**Vibration, Acceleration, Pressure breathing, exercise**
1	**Physiological**	**Illness, Sleep deprivation, Dehydration, Fatigue, Alcohol, Narcotics**
2	**perceptual**	**Poor communication channel, Poor grasp of the language used, Noise (Lombard effect)**
3	**psychological**	**Emotion, workload, Task-related anxiety, Background anxiety**

2.5 Measuring stress

Availability of modern sensor technologies has enabled realistic measurement of stress in humans (Bakker et al., 2012). It enables measurements of variables such as heart rate, galvanic skin response and facial expression. A strong correlation was found between these variable and stress level (Bakker et al., 2012). In measuring physiological induced stress sensor based approaches are mostly employed. Sensors based stress measuring techniques include chemical analysis, skin conductance, and electrocardiograms etc. These techniques however requires bodily contact with their user (Dhole & Gurjar, 2013). Apart from being costly, cumbersome and inconvenient, these approaches are likely to be a source of stress on the user. This has motivated seeking alternative approach that is cheap, easy to use, convenient, and not stressful. In attaining this objective, Lu et al., (2012) proposes smart

phones approach for ubiquitous and unobtrusive stress detection. In the same vain, Bakker et al., (2012) applies associative classification approach in data mining to detect, predict, and provide training on stress management. However of this approaches, speech signal approach to stress detection comes in handy, convenient, cheap, effective, and free from stress inducement.

2.6 Stressed speech features

Stress has varying effect on speech production and consequently on the acoustic signals. The theme of many studies in literatures centres on stress and its realization in the acoustic signal(Narayana & Kopparapu, 2009).Though the effect of stress follows the same pattern on speech production and acoustic signal, however, the magnitude of the effects differs across speakers and acoustic features (Bou-Ghazale & Hansen, 2000). Efforts were made by several researchers to identified most suitable stressed speech features. In detecting stress from speech, several spectral and acoustic features are being employed. Features such as fundamental frequency (pitch), spectral energy, amplitude, duration and several others were commonly used (Lai, Chen, Chu, Zhao, & Hu, 2006). The fundamental frequency of voice (*f0*) is the most widely considered spectral feature in voiced speech, mainly with vowels (Sigmund, 2010).The most widely extracted features from stressed speech include fundamental frequency (*f0*), formants, amplitude, intensity, duration, and rate of speech (ROS) (Hong, Ramos, & Dey, 2012).

Stress analysis experiment is usually carried out through analysis of some stress parameters such as fundamental frequency (*f0*), pitch, vowel duration and formants from the

recorded emotional speech, obtained under stressor conditions like stress, environmental noise, fatigue, heavy workload, and/or loss of sleep (Lai et al., 2006).

Stress researchers classified speech as (a) uttered when the speaker is under stress or expressing some emotion and (b) uttered in a neutral speaking style, namely, when the psychological state of the speaker does not seriously affect the speech, for example, reading news, watching TV, or even normal conversations.

It is evident from the literatures that most researchers consider speech of type (a) when conducting experiment on stressed speech. Contrary to this assertion, Narayana and Kopparapu (2009) observed that speech either be it emotional or normative has an element of stress in some syllables of speech, which to a larger extent is determined by some influencing factors such as language, accent and the geographical location from which a speaker originated from. This they argued further that the corresponding pitch and amplitude contours would have been completely smooth if the stress is in not present in non-emotional speech. According to them what distinguishes stressed or non-stressed speech is the magnitude of the stress parameters and that in both cases, the stress parameters are the same.

In stress detection related researches, the most widely used acoustic cues for stressed speech are higher (*f0*), greater duration and higher intensity. Of equal note is the fact that there is strong correlation between stressed speech and voice quality (Narayana & Kopparapu, 2009).

2.7 Speech corpus

Lack of appropriate stressed speech corpora has constituted a major obstacle towards study of stressed speech (Scherer *et al.*, 2008; Sigmund, 2010). Again the ability to obtain accurate voice samples of speakers in different situations of stress, recorded in real conditions has remained an uphill task. Likewise, simulation of real situations stressed speech by normal speakers or actors have not yielded the desired perfect real case stressed corpus. According to Sigmud (2010) there are some methods to simulate stressful events such as using vocal noises, quick question-answer-quizzes with the opportunity to win a prize, negotiation regarding an important contract, etc.

There are several databases that are been used in stress detection experiment (Hansen et al., 2012; Schuller, Vlasenko, Eyben, Rigoll, & Wendemuth, 2009). These include SUSAS (*Speech Under Simulated and Actual Stress)*, EMO-DB (Berlin Emotional Speech-Database), ATCOSIM (Air Traffic Control Simulation Speech corpus) etc. The most widely used and reported corpus in the literature is the SUSAS. This database, called *SUSAS,* has been employed extensively in the study of how speech production and recognition varies when speaking under stressed conditions.

The SUSAS database contains speech under simulated and actual stress grouped into five different categories that were developed specially for speech under stress investigations. The vocabulary covers 35 single-word utterances from aircraft communication. This absence of a general and reliable stressed speech corpus (Wang, 2009) has motivated researchers to collect a more constrained but more reliable stressed speech corpus.

2.8 Speech recognition

ASR is an approach for converting speech signal into spoken word equivalents in text or command for executing action (Deng & Li, 2013). Based on Rabiner et al. (1978), Speech-To-Text (STT) i.e. ASR is the process of converting or translating an acoustic signal, captured over a microphone or a telephone, and mapping it into a set of words. A word is recognized by extracting features from captured signal and classifies the features based on the given voice sample in the database (Adami, Lazzarotto, Foppa, & Couto Barone, 1999). With pervasiveness of technology, speech is far becoming a preferred approach for control and command between humans and machines (HCI) (Gray, 2006; Ibiyemi & Akintola, 2012) and has thus found adoption in several applications.

Among the wider applications of stress detection includes but not limited to the area of crime detection or countermeasures (e.g., kidnapping cases, lie detector systems, analysis of suicide cases involving phone calls), safety and security (e.g., involving pilots and air traffic controllers in highly stressful noisy environments, or drivers involves in heavy traffic jams) and psychology/health (e.g., monitoring stress and emotional level of patients). Though intensities of stress vary severely based on situations, analyzing the content of speech signal under stress is a very important factor in speech analysis and recognition (Wang, 2009).

The goal of ASR is to attain performance comparable to unconstrained human speech perception. However, unconstrained ASR has remained a difficult task. Hence most ASR systems achieved a reasonable performance by artificially constraining the problem (Acero, 1990) such as phonemes (vowels) recognition (M. M. Azmi & Tolba, 2008; Mohd Yusof & Yaacob, 2008; Siraj, Shahrul Azmi, Paulraj, & Yaacob, 2009), word recognition, sentences or

limited vocabulary. Phonemes are the smallest unit of word generation and hence in natural languages, phonemes are crucial in formulating word meaning (Siraj et al., 2009). Because in generating words of a language, Consonants-Vowel (CV) units which has highest occurrence, Azmi &Tolba (2010) argues that recognition accuracy of ASR can be improved by increasing the ability of ASR to recognize vowel in CV unit.

2.8.1 Feature extraction

Feature extraction is the process of extracting unique characteristic that provides a compact representation of the given speech signal and the output of this process are vectors coefficients (Acero, 1990). An effective and efficient feature extraction technique largely determines the accuracy of ASR system (Sun, Yuan, Bebis, & Louis, 2002). The fact that the success of pattern classification depends on an accurate feature extraction has made feature extraction the most important component of ASR (Hamdy, Hefny, Salama, Hassanien, & Kim, 2012).

The feature extraction involves three stages. Spectra-temporal analysis of the signal and generation of raw features that describes the envelope of the power spectrum of short speech intervals is performed in the first stage. Compilation of an extended feature vector made up of static and dynamic features is carried out in the second stage. Lastly in the last stage, transformation of extended feature vectors into more compact and robust vectors that are then delivered to the recognizer (Anusuya & Katti, 2010).

Though there are several speech features extraction techniques such as Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP) and Mel-Frequency Cepstral Coefficient (MFCC), however, MFCC has remain the most widely used method particularly

in speech recognition and likewise speaker verification applications (Abu Shariah, Ainon, Zainuddin, & Khalifa, 2007).

Several speech features for speaker recognition and speech recognition applications have been proposed in literature. Mel Frequency Cepstral Coefficients (MFCC) by far, have been the most commonly used speech features (Narayana & Kopparapu, 2009). Mel-Frequency Cepstral Coefficients (MFCC) is the most dominant and foremost method for extracting spectral features inherent in speech signal. MFCCs based on frequency domain using the Mel scale that was fashioned based on the human ear scale (Razak, Ibrahim, & Idna Idris, 2008), because MFCC take into consideration the characteristics of the human auditory system, it is commonly used in the ASR.

MFCCs being considered as frequency domain features are much more accurate than time domain features. According to Huang et al. (2001) MFCC denotes the real cepstral of a windowed short-time signal obtained by performing Fast Fourier Transform (FFT) of speech signal.

Extraction of MFCC features employs a frame-based analysis of a speech signal in which the speech signal is broken down into a series of frames. Sinusoidal transform (Fast Fourier Transform) is performed on each of the frame so as to find definite parameters that are further subjected to Mel-scale perceptual weighting and de-correlation. The output of this process is a series of feature vectors defining simplified frequency information and useful logarithmically compressed amplitude (Buchanan, 2005). In computing MFCC, the following steps are involved: (i) pre-processing, (ii) framing,(iii) windowing through hamming window, (iv) Discrete Fourier Transform (DFT), (v)performing mel-scale filter bank in order to find

the spectrum as it might be perceived by the human auditory system, (vi) taking both the Logarithm, and the inverse DFT of the logarithm of the magnitude spectrum (Daniel & Martin, 2009; Khalifa, El-Darymli, Abdullah, & Daoud, 2013). MFCC feature extraction process is as shown in figure 2.2 below.

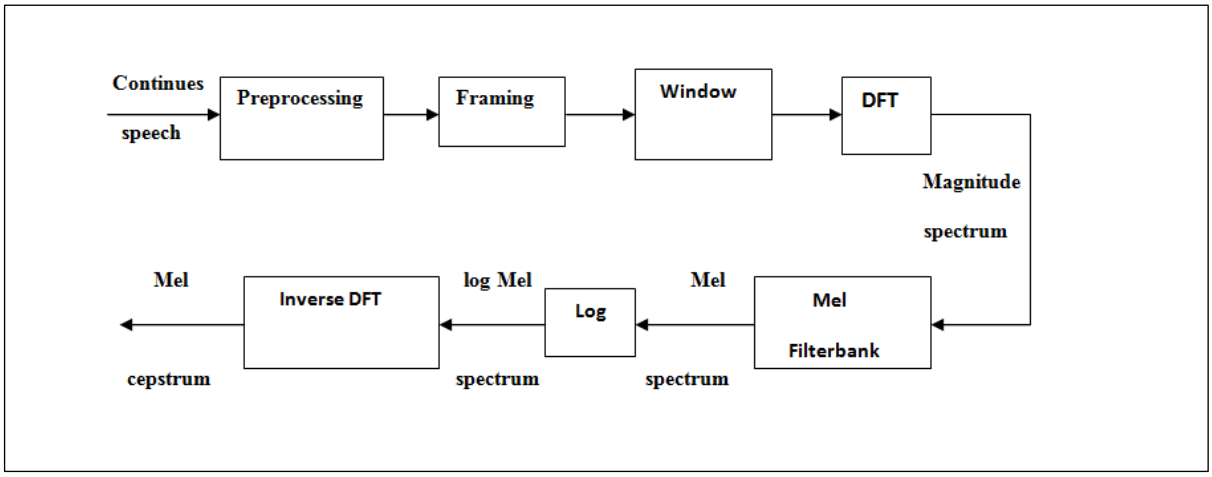


Figure 2.2. MFCC Feature Extraction Process.

2.8.2 Pattern recognition

Speech recognition problem fits into a much wider scientific theme known as pattern recognition or pattern matching/classification. According to Huang et al. (2001), the heavy reliance of spoken language processing on pattern recognition, has been one of the most challenging problems for machines. It is a process by which extracted features/patters is been matched or assigned to one of the many arranged classes (Haykin, 2009).

Pattern recognition/matching involve weighing the resemblance between two speech patterns: the unknown speech to be recognised and known that was obtained from the training process of each element that can be recognized (Madisetti & Williams, 1999). Pattern recognition is aimed at classifying objects of interest into one of a number of categories or classes(Anusuya & Katti, 2011). The output of feature extraction process i.e. sequences of

acoustic vectors which are usually called patters becomes the object of interest in the recognition stage. Based on the nature speech corpus, classes can refer to individual words, sentences, or phonemes. Madisetti and Williams (1999) as shown in figure 2.3 explained that the pattern recognition model is made up of some elements which include speech analysis, pattern training, pattern matching, decision strategy and templates or models containing the pattern training features for pattern matching purposes.

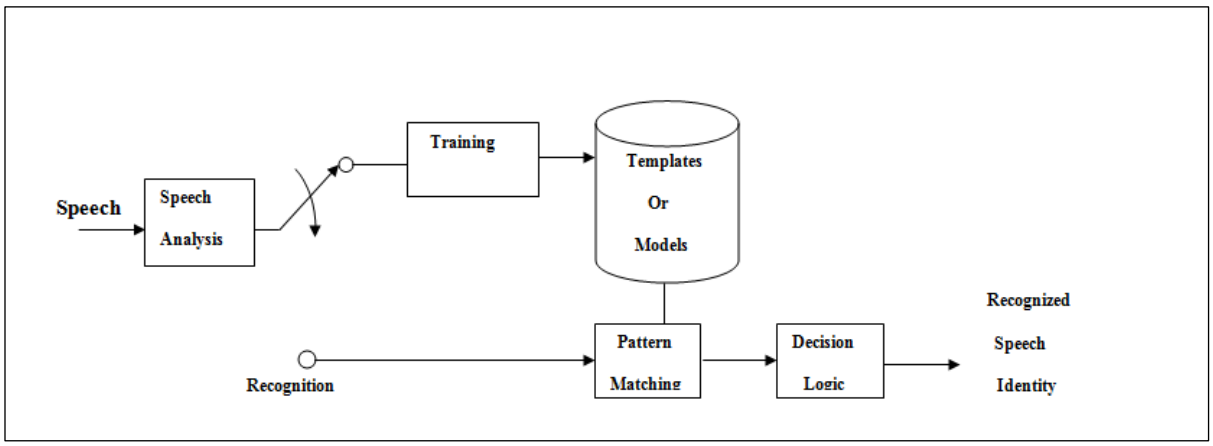


Figure 2.3. Block Diagram of Pattern Recognition Speech Recognizer (Madisetti and Williams, 1999)

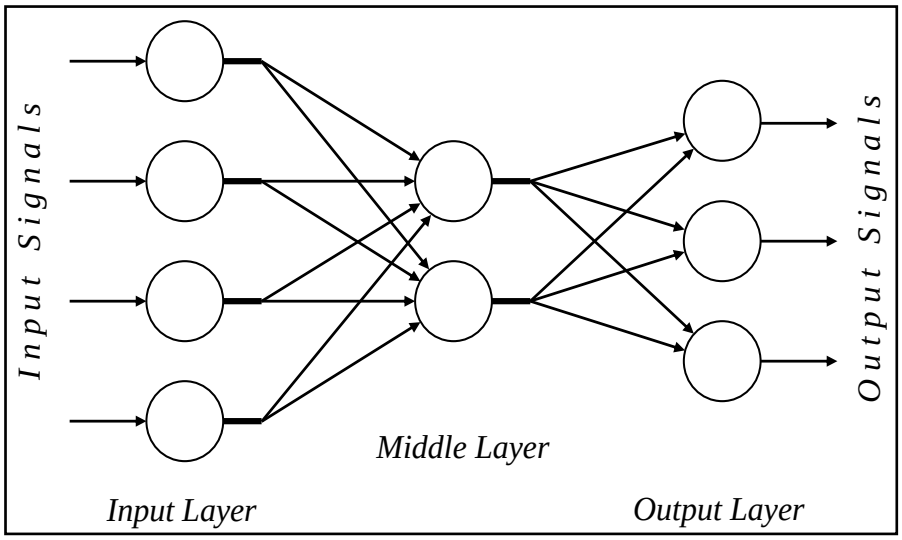
Several methods of pattern classification includes Hidden Markov Model (HMM), Neural Networks (NN) and Vector Quantization (VQ), Bayesian Networks (BN), Support Vector Machine (SVM) ,KNN, LDA, LM and much more (Deng & Li, 2013). Of these methods of the most used recently is NN. Neural Networks (NN) or Artificial Neural Networks (ANN) is an artificial intelligence approaches aimed at mechanizing the recognition process likened to the manner by which human uses intelligence in visualizing, analysing, and characterizing speech from set of acoustic features (Madisetti & Williams, 1999). Inspired by biological models of the nervous system, Neural Networks' structure is structured as a model of the human brain's activities aimed at imitating certain processing capabilities of the human brain. The neural network is built-up of many artificial neurons, known as perceptrons.

2.8.3 Artificial neural network as vowel classifier

Multi-layer Perceptron (MLP) was used as an artificial neural network. Multi-layer Perceptron is used to classify the Malay vowels. An artificial neural network consists of a number of very simple processors, also called neurons, which are analogous to the biological neurons in the brain. The neurons are connected by weighted links passing signals from one neuron to another. The output signal is transmitted through the neuron's outgoing connection. The outgoing connection splits into a number of branches that transmit the same signal. The outgoing branches terminate at the incoming connections of other neurons in the network.

Neural network architecture is made up of several layers of nodes as shown in Figure 2.4. They are an input layer, an output layer and zero or more hidden layers. Each independent variable is represented by a node in the input layer, while the final decision or target is represented by one or more node. The number associated with each link between nodes is called a weight. Feed-forward networks can feed their outputs in only forward direction.

A Multilayer Perceptron (MLP) is a feed forward neural network with one or more hidden layers was used in this research to identify the vowel utterances. A model is built describing a predetermined set of data classes or concepts and used for classification. The words "CAT, BED, SIT, BOW, CUE" are represented by the 3-bit output neurons. The network was trained and tested using 70% and 30% of the data respectively using learning rate of 0.1 and momentum factor of 0.9. The weights and biases of the network were initialized randomly. The percentage of test set samples that are correctly classified, then hold out, k-fold cross validation method from fold 1 up to fold 10. The accuracy of the model is based on the training set and test set.



Fig*ure 2.4.* Architecture of a typical artificial neural network

A Levenberg-Marquardt (LM) network with a single hidden layer was used in this research to identify the vowel utterances. There are 4 to 14 input neurons for features extracted from different feature extraction methods and 1 output neurons for each targeted vowels. 3 The network was trained using 70% of the data using learning rate of 0.1 and momentum factor of 0.9. The weights and biases of the network were initialized randomly. The output class that will be activated will be based on highest of the 6 output neuron values.

2.8.4 Linear Discriminant Analysis

Discriminant analysis is a statistical technique to classify objects into mutually exclusive and exhaustive groups based on a set of measurable object's features (Chee et al., 2009). Term discriminant analysis comes with many different names for different field of study. It is also often called as pattern recognition, supervised learning, or supervised classification.

In discriminant analysis, the dependent variable (Y) is a group and the independent variables (X) are the object features that might describe the group. The dependent variable is always category (nominal scale) variable while the independent variables can be of any measurement scale (i.e. nominal, ordinal, interval or ratio). If the groups are assumed to be linearly separable, we can use Linear Discriminant Analysis model (LDA). Linearly separable suggests that the groups can be separated by a linear combination of features that describe the objects. If there are only two features, the separators between objects group will become lines. If the features are three, the separator is a plane and when the number of features is more than 3, the separators become a hyper-plane.

2.8.5 K-Nearest Neighbors

A supervised classification classifies an unlabelled object into a suitable class using some labelled objects. An unsupervised classification does not use the labelled objects. KNN is well known as a basic method of supervised classification (Duda, Hart, & Stork, 2001). It is a simple algorithm which is based on the assumption that the examples residing closer in the instance space have same class values. Usually Euclidean distance is used as the distance metric which works with numerical values.

K-nearest neighbor is a supervised learning algorithm where the result of new instance query is classified based on majority of K-nearest neighbor category. The purpose of this algorithm is to classify a new object based on attributes and training samples. Given a query point, we find K number of objects or (training points) closest to the query point. The classification employs majority vote among the classification of the K objects. It uses neighbourhood classification as the prediction value of the new query instance.

2.9 Stress Classification Related Works

The focus of the past work on short-term speech variation has been on speech features extraction and analysis, the design of detection systems, and the development of robust of speech ASR and on stress detection based on speech signal analysis.

The possibility of detecting alcoholic consumption level in human through speech processing was experimented by (Levit, Huber, Batliner, & Noeth, 2001). The speech corpora used was made up of alcoholized speech samples collected at the Police Academy of Hessen, in Germany. It contains 120 readings utterances by 33 male speakers in different alcoholization conditions with alcohol blood level ranging between 0 and 2.4 per mile. Extracted features include *f0,* energy, and duration of voiced and unvoiced intervals, MFCC, and jitter. Resultant features were trained and tested by using Multi-Layer Perceptron (MLP) of Artificial Neural Networks (ANN) as classifier. Recognition rate of 80.2% was achieved using a combination *f0*, energy, and jitters features.

Zhou, et al., (2001) augured that there is inconsistency on the part of most of the features being examined by researchers as indicators of stress. Three features of frequency modulation (FM), autocorrelation envelope area and critical band (CB) were derived from Teager energy operator (TEO) based on the assumption that features derived from TEO gives better indication of stressed speech. Simulated stress from "talking styles" and actual stress from "amusement park roller-coaster" domains from SUSAS speech corpus were used for the assessment. The recognition task was based on the subset of SUSAS words: "freeze," "help," "mark," "nav," "oh," and "zero." Classification performance result from HMM of TEO and MFCC features reveals that TEO is more consistent with stress condition and hence substantially outperforms MFCC.

The difficulty of implementing automatic stress recognition system due to lack of reliable stressed dataset motivated the work by Scherer et al., (2008). Experiment was performed with fifteen subjects that were subjected to air traffic controller simulation game to induce stress in the participant. In performing the experiment, the subjects are required use the mouse to select and change the flight routes of several planes flying at different speeds representing difficulty levels, with the aim of avoiding collision, while simultaneously say the correct answers for the questions shown on top of the screen. Modulation spectrum features was extracted from the corpus formed by the speakers responses to the questions. These features serve as input to the Recurrent Neural Networks (RNN) classifier. The result shows a better performance of RNN compare to those obtained by human classifiers.

Gaussian mixture model based framework for detecting physical stress induced by execrising on the physical task of using a stair-stepper for 10 minutes at a constant speed of between 9-11 miles per hour was proposed by (Patil & Hansen, 2008). The corpus is made of 35 speech utteraces recoreded from each of the 42 female speakers while in neutral and prompted physical stress conditions. A combination of mel-scale cepstral coefficients and Teager energy operator was obtained by using Adaboost. Classification accuracy of 73% classification accuracy with a generic stress model was realised.

The fact that speech contains large amount of speaker information was exploited by (Narayana & Kopparapu, 2009) to improve performance of speaker recognition system by using stressed speech information. In conducting their experiment, 6 English words repeated severally were recorded from 4 speakers. Pitch (*f0*), amplitude, and duration (PAD) values were extracted from the recorded speech. The mean, variance, and standard deviation were calculated from PADs for each speaker. The minimum Euclidean distance between the

training and the test data was use as basis of classification. Classification results reveal that PADs are good features of stressed speech.

Wang (2009) researched on speech recognition under stress condition based on word recognition. Prosodic features of pitch, intensity and glottal spectrum together with LPC features was employed in the research. Speech corpus used is SUSAS while three different classifiers of DTW, HMM and Hidden Control Neural Network (HCNN). Results shows that pitch, intensity, and glottal spectrum are all affected by stress. Of the three classifiers, HCNN has better recognition performance of 90%.

Sigmund (2010) experimented on the effect of oral exam stress on fundamental frequency (*f0*) and spectrum to measure the possibility of speech detection through speech. ExamStress speech corpus was populated with data collected from 31 male students prior to and during final oral exams uttering spontaneous speech in Czech. Heart rate readings were taking during the first 900s of both pre and during oral exam. Mean and standard deviation were calculated for both neural and stressed speech values extracted by autocorrelation. The result shows increase in average and range during stressful situations. While the spectrum shows similar results, it however reveals that vowels “e” and “i” are mostly affected by stress situations. This finding is of significant in stress detection as sharp changes in the values of vowels “e” and “i” is an indication of stress.

The need to unobtrusively and ubiquitously detect and monitor stressin real-life situation motivated the development of StressSense by Lu, et al. (2012). StressSence involves detection of stress using smartphone microphone and adapting it to users by Maximum A Posteriori (MAP) adaptation. Experiment speech data were collected form 14 speakers made

up of 10 females and 4 males. Data collectetion was segmented into 3 groups of: neutral reading, job interview and marketing tasks that represents stressed situation. The recording was done with Google Nexus smartphone. Pitch, spectral centroid, ROS, MFCCs and TEO-CB features extracted from speech corpus was feed into GMM for classification. Their experimental result reveals that pitch and ROS are the most stress affected features.

Stress@work framework developed by Bakker et al., (2012) aimed at measuring and predicting stress level, and providing training on stress management at workplace. Physiological data were obtained with the aid of GSR and accelerometer from 5 employees for period of seven weeks. On the daily basis, each participant downloads their respective data from the sensor devices into their computers. This is followed by tagging their daily activities depending on its stress level as scheduled in MS outlook. Visual exploration of collected data was done at the end of the seventh day. GSR readings were plotted and compare with MS outlook work schedule of the participants. Though the result of their experiment is encouraging, it is however being marred by challenge of getting accuracy and consistence reading from the GSR. They nonetheless admitted the need to combine their approaches with others such as speech and facial data.

In the work presented by Dhole et al. (2013), spectral analysis of the speech signal was performed. Speech duration and amplitude were extracted as speech features for classification purpose. ANN was used as classifier. Frame level feature extraction was performed in the study. Speech samples were taken from the viva contestant aged between 22 to 24. For stress analysis, speech sample were taken before and after the viva session. All the collected samples were used as dataset for stress in speech using spectrum analysis of the speech. Only low frequency component of the speech was considered into account for classification.

According to Dhole et al. (2013), low frequency components provide better characterization than high frequency components. Discrete Wavelet Transform (DWT) was used for converting speech signal into low frequency by using digital filtering technique. An average of 84.72% result was reported for conducted experiment.

He et al. (2009) conducted study of stress at vowel level. Spectrograms features were calculated for classification. The whole speech utterance was divided into 256 points frames with 196 points overlapping each other. Spectrogram show increase in formant energy and also increase in pitch for strong stress. This indicates that spectrogram can be used to differentiate between different levels of stress. He et al. (2009) have used two classifiers GMM and KNN for classification process. SUSAS database was used as input speech for training and testing the classifiers. Speech signals under actual stressful working conditions were used as datasets for classification of three classes: high level stress, low level stress and neutral. Three datasets A, B and C were created that represent vowels at each class. Dataset A contains words with different vowels, dataset B contains words with single vowel ‘a’ and dataset C contain words with single vowel ‘e’. For each datasets, 80% data was used for training and 20% data was used for testing. For each dataset (A, B and C), classification with GMM and KNN was performed 15 times. An average classification rate was calculated over 15 runs for each dataset and classifier combination. Result shows that both classifier GMM and KNN gave similar results. Proposed model has given accuracy ranged from 55.34% to 73.76% depending on the chosen dataset and classifier. Different datasets also given very small difference in accuracy i.e from 55% to 70% for set A, 57% to 73% for B and 57% to 72% for C. This indicates that vowels have very small effect on accuracy.

Frame level feature extraction was performed by Chee et al. (2009). MFCC speech features were chosen for classification purpose. Two classifiers, KNN and LDA were used for classification. Speech utterances were taken from University College London Archive of stuttered Speech (UCLASS). UCLASS consist on 43 different speakers with 107 reading recordings. For experimental purpose, 10 samples of speech from UCLASS were taken. Recordings of 2 female readers and 8 male readers for one each sample was taken. Chee et al. (2009) stated that 90% accuracy can be achieved while using MFCC with KNN and LDA. Different values of 'k' in KNN classifier were also studied in the experiment. It is said that lower values of 'k' provide better accuracy than higher values of 'k'. Chee et al. (2009) mentioned that 'k' values ranged from 1 to 2 provide high accuracy than values ranged from 3 to 10. Summary of related research studies is as presented in table 2.2.

Table 2.2

Summary of related research studies

Researcher(s)	**Project**	**Corpus**	**Features**	**Feature extraction**	**Pattern classification**
(Levit et al., 2001)	Alcoholized speech	120utterances, 33 males	*f0*,energy, duration	MFCC	MLP (ANN)
(Zhou et al., 2001)	Stress detection	Amusement park coaster-roller (SUSA)	FM,auto-correlation envelope CB form TEO	MFCC	HMM
(Scherer et al., 2008)	Mental stress detection	15 subjects	Modulation spectrum features	-	RNN
(Patil & Hansen, 2008)	Physical task stress detection	42 females speakers, 35 utterances each	TEO	MFCC	GMM

(Narayana&Kopparapu, 2009)	Stressed speech for speaker recognition	4speakes	Pitch,amp -litude & duration (PAD)	-	Euclidean distance
(Wang, 2009)	Stress recognition is speech	SUSAS	Pitch, intensity, &glottal spectrum	LPC	DTW, HMM, & HCNN
(Sigmund, 2010)	Exam Stress	31 males	*f0*,spectru m & heart rate	Autocorrel-ation	-
(Lu et al., 2012)	ExamStress detection with smartphone	14 speakers (10 females, 4 males)	Pitch, spectral centroid, ROS, TEO-CB	MFCC	GMM
(Bakker et al., 2012)	stress@work	5 participa-nts	Heart rate & GSR	-	-
(Dhole et al., 2013)	Detection of speech under stress	Viva contestant	duration and amplitude	-	ANN
(He et al., 2009)	Stress detection using Spectrogram features	8 male 2 female	-	-	KNN,GMM
(Chee et al., 2009)	MFCC based recognition of stress	43 speakers	-	MFCC	KNN,LDA

CHAPTER THREE

METHODOLOGY

3.0 Introduction

Research approach to be used in this work discussed in this section. It consists of research design, databases, experimental setup and process undertaking to achieve the dissertation objectives.

3.1 Research Design

Experimental research approach is used for this research work. This research approach is adopted because of the exploratory nature of this study, and its appropriateness to the research questions. Research Framework is shown in Figure 3.1. The framework consist of data collection, pre-processing followed by features extraction of both prosodic and MFCC. Pattern classification is done using three classifiers of KNN, LM and LDA consisting of training with 70% data and testing with 30%. The speech features is classified as either neutral or stressed speech. If stressed, it is further classified as low, medium or high. The details of the research framework are as given in sections 3.2 through 3.8.

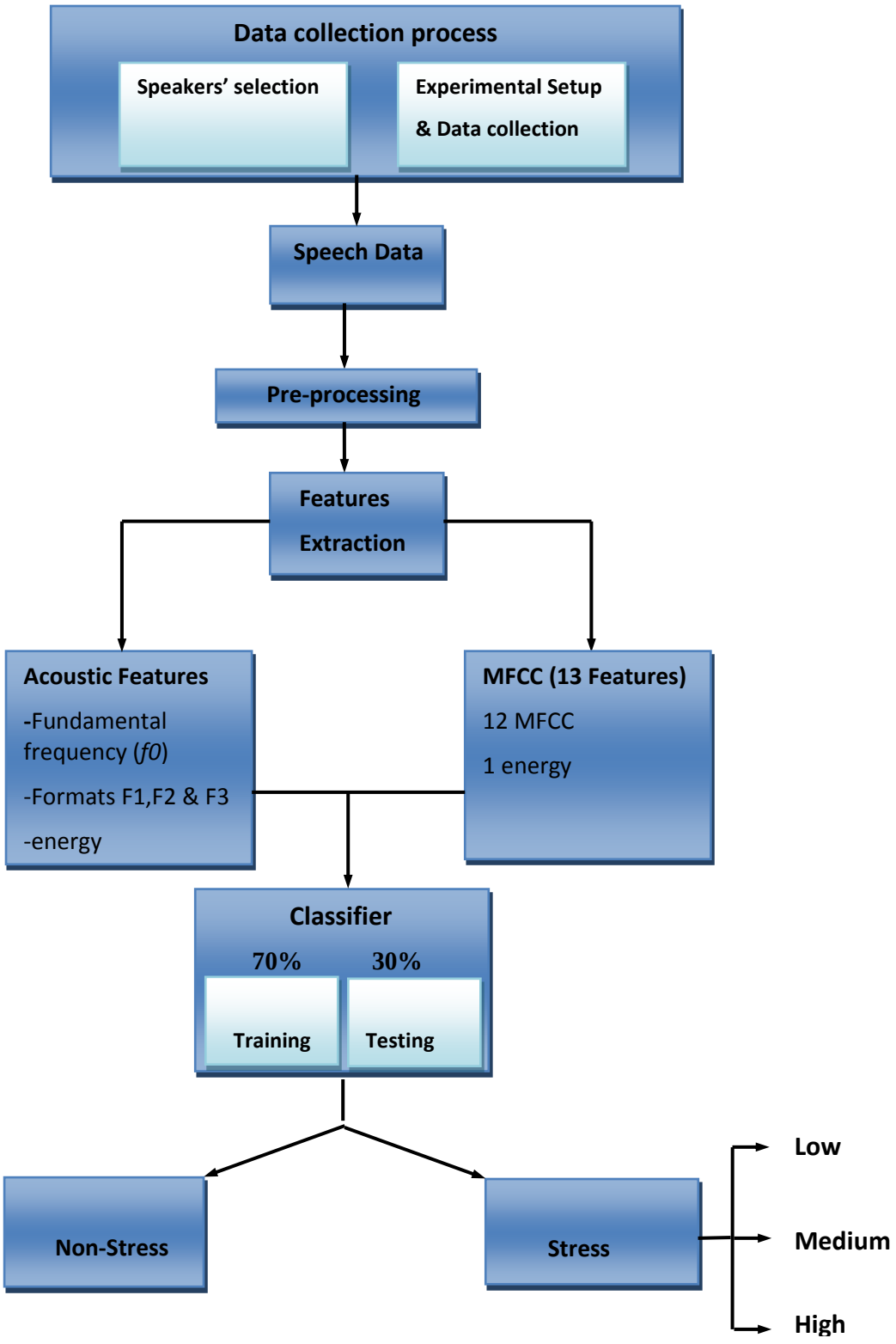


Figure 3.1. Research Framework

3.2 Speakers' selection

A total of 10 speakers made up of 7 male and 3 female both of Arab origin who are students of UUM were selected for data collection by stratified random sampling approach. Selected speakers are those that have not been exercising regularly to guarantee that the speakers are exposed to the same exertion condition. To ensure that the measured stress is actually induced by the cycling exercise, the recording is to be done in the morning when the speakers would not have been exposed to any form of stress. In the similar vein, in the night prior to the speech recording the speakers were encourage to sleep early so as to mitigate the effect of stress as a result of inadequate sleep.

3.3 Experimental setup

The experiment process was divided into two segments based on type of speech to be collected. The speech in the corpus is categorized into neutral and stressed speech. Neutral speech represents utterances made by the speakers when in a relaxed mood. It is taken while the speaker is in a sitting position. The stressed speech on the other hand was obtained while the speaker is cycling maintaining a speed of about 10 mph on the Free Motion exercise bicycle in the gym.The recordings is done by using high definition head microphone and a laptop computer using Matlab scripts and a sampling frequency of 8 KHz. A Sound Level Meter (SLM) is used to measure the gym environment to ensure that the Sound to Noise Ratio is within acceptable limit of about 40db. To represent the five English vowels: /a/, /e/, /i/, /o/ and /u/, the words "CAT, BED, SIT, BOW, CUE" were used to represent the five vowels for the fact that energy level in vowels far outweighs that present in consonants (Mohd Yusof & Yaacob, 2008). It is evident from human physiology literatures that several physiological variables such as heart rate, skin resistance, and blood volume are significally

affected by the sympathetic unit of human automatic nervous system (ANS) (Zhai, Barreto, Chin, & Li, 2005). Hence heart rate meaurement will be part of the experiment.

Prior to the start of the recordings, a heart rate monitor watch is affixed to the speakers to measure the heart rate. Heart rate readings are taking at intervals of 15s. In each set of the recordings, the words "CAT, BED, SIT, BOW, CUE" are to be repeated 5 times by each speaker. The recordings were to be repeated five times per speaker to ensure sufficient data sets. Table 3.1 shows details of the experimental setup procedure.

Table 3.1
Experimental setup details

Corpus details	***Neutral Speech***	***Stressed Speech***
Speakers	10 UUM students (7 Male & 3 Female)	10 UUM students (7 Male & 3 Female)
No. of utterances	250 (5x5x10)	250 (5x5x10)
Sampling frequency	8khz	8khz
Uttered words	cat, bed, sit, bow, cue	cat, bed, sit, bow, cue
Heart Rate measurement	15s intervals	15s intervals

3.4 Data collection

The process of data collection from the experimental set up is represented in the Figure 3.3 below. The process starts with the selection of speakers from whom the data is to be taken. To ensure high quality of the recorded data, the mouth piece of table top microphone is well placed below the lips of the speakers. The speakers were then briefed on the objectives of the experiment and the modalities for data collection. Before actual data were collected, speakers undergoes few runs of rehearsal to get them familiarize with the process. The actual recording is grouped into two session of neutral taken while the speakers are in relaxed mood and stressed taken while the speakers is actually exercising on the stationary bicycle. Matlab

program was used to capture the speakers' utterances. The quality and the numbers of the recorded utterances is checked and saved as wav file to the laptop is satisfactory. If not, the recording is discarded. This process is repeated until data has been collected from all the selected speakers.

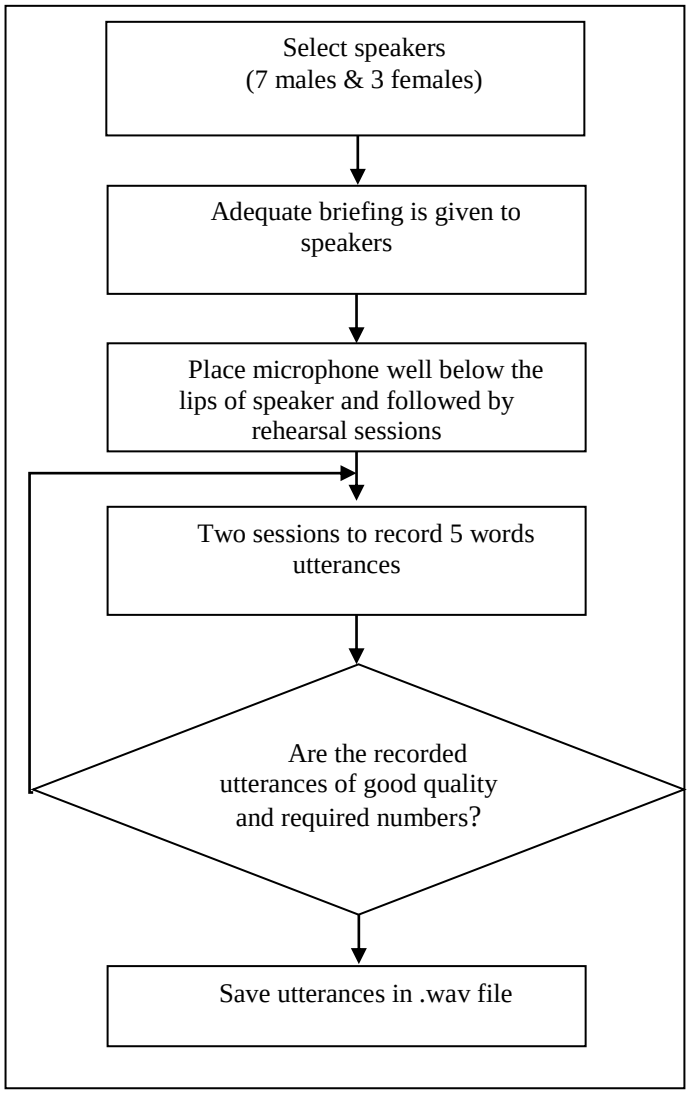


Figure 3.2. Data Acquisition Process (Adapted from Azmi, 2010).

3.5 Speech corpus

The success of any experimental research most especially speech recognition to a large extent depend on the availability of accurate data. This is because the validity of the results of any experiments depends on it. Unlike emotion data where there are divergent

opinions among researchers on the reliability of laboratory simulated emotion, data collection for physical stress is fortunate to avoids much of the ambiguity associated with the collection emotion data (Godin et al., 2009). The corpus for this research code named UUM-CycleStress was collected from selected Arab students of the Universiti Utara Malaysia. The corpus is made up of five English words: “CAT, BED, SIT, BOW, and CUE” repeated five times each spoken by 7 males and 3 females both in neutral and stressed conditions making a total of 500 utterances. Sequel to feature extraction, the recorded speech is subjected to pre-processing process to clean up the speech signal of unwanted components such as noise.

3.6 Pre-Processing

To ensure the quality of speech signal, pre-processing stage is carried out prior to feature extraction. Pre-processing involves series of steps to normalize the characteristics of the speech signals recorded in the time domain. stages of pre-processing includes: (i) speech segmentation/ end-point detection of the starting and final speech signal point, (ii) normalization of the recorded speech signal (iii) pre-emphasize and (iv) windowing.

3.6.1 Segmentation/ Endpoint Detection

In detecting the beginning and final point of a word, the energy activity, and zero crossing rate of the speech signal, $S(n)$, in considered in relative to the values of silence a state. The start and endpoint locations for word segmentation were done considering energy method and zero crossing rates.

3.6.2 Normalization

In other to minimize the effect of variations of the amplitude in the course of speech

Recording process due to speaker position with respect to the microphone, fatigue, or distraction, amplitude normalization of speech signal was conducted in order to obtain a greater similarity between files that contain the same word.

3.6.3 Pre-Emphasizing

In speech processing, a process called pre-emphasis is applied to the input signal before the LPC analysis which can be reversed with a de-emphasis process during the reconstruction following the LPC analysis. The pre-emphasis is used to compensate the lost that suffer the high speech signal frequencies by effect of the propagation and radiation from the vocal cavity to the microphone. As the frequency increases, pre-emphasis raise the energy of the speech signal by an increasing amount. The pre-emphasis is performed by filtering the speech signal with a first order filter whose output signal is given by passing the speech signal through a first order filter. This filter improves the efficiency of the stages used to calculate the speech spectrum, increasing, from the hearing point of view, the sensibility of the frequencies components larger than 1 KHz.

The pre-emphasize filter was implemented by using a pre-emphasized constant value of 0.95 where *s**(*n*) is the pre-emphasized signal.

$$s^*(n) = s(n) - A_p s^*(n-1) \quad (3.1)$$

Where

s(n)* – pre-emphasized signal
s(n) - original signal
A_p – pre-emphasize constant (0.95)

3.6.4 Windowing

Window is a finite length sequence used to select a desired frame of the original signal by a simple multiplication process. Some of the commonly used window sequences are rectangular, Hamming, Hamming and Blackman, and so on. Usually a Hamming window is used which is given by equation (3.2). Windowing in time domain results in a convolution in the frequency domain of the signal spectrum and in the window spectrum. In this research, hamming window was chosen because it is the most widely used by speech recognition researchers.

$$w_H(m) = \begin{cases} 0.54 + 0.46\cos(\pi m / M) & -M \le m \le M \\ 0 & otherwise \end{cases} \quad (3.2)$$

3.7 Feature extraction

Three acoustic features of fundamental frequency (*f0*), formant (F1, F2 and F3), and energy were extracted using Wavesufer (Sjölander & Beskow, 2000). Spectral feature based on MFCC are extracted from both the neural and the stressed speech. The mean, range and standard deviation of *f0*, F1, F2, F3 and energy will be calculated. These calculated values are to be plotted and comparison between neutral and stressed values is done. This is for the purpose of determining the effect of physical stress of speech production and to gain further insight into the features that is mostly affected by stress.

3.8 Pattern classification

The extracted acoustic features and MFCCs served as input to learning systems, which were trained to identify the stress speech from the normal speech of a person undergoing

physical exercise. Three popular learning algorithms were employed: Levenberg-Marquardt (LM) network, k-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA) for both the learning and classification process. Matlab software that is made up of collections of machine learning algorithms for classification tasks is used to implement all three learning systems (Mohd Yusof & Yaacob, 2008).

Classification system is built for each speech types (neutral and stressed) and combination of both speech types. And also for the acoustic features and MFCCs. Different combinations of the features were feed into the classifier to determine features that yielded better classification result.

In determining these features, the features combinations were varied until the best overall world error rate (WER) was found or higher classification rate (CRs) is attained. The classification process is made up of two tasks of training and testing.

.

3.8.1 Training

Based on similar researches (Deng & Li, 2013; Mohd Yusof & Yaacob, 2008), 70% of the speech corpus will be used in training the three classifiers – KNN, LM and LDA. Three sets of acoustic model will be trained with 3 sets of neutral data, stressed data, and both. The input features into the classifiers consist of two groups of:

(i) three acoustic features of
   a. Fundamental frequency (*f0*)
   b. First, Second, and Third formants (F1, F2 and F3)
   c. Energy

(ii) Mel Frequency Cepstral Coefficients (MFCC).

3.8.2 Testing

In determining the accuracy of the classifier, 30% of the speech corpus will be used for evaluation (M. Y. S. Azmi, Idayu, Roshidi, Yaakob, & Yaacob, 2012). The performance of the classifier will be based on percentage of correctly recognized (CRs) vowels under both the neutral and stressed conditions.

3.9 Stressor

Based on earlier assumption, the presence of a stressor implies that the speech is stressed. The stressor in this study was a stationary bicycle (exercise bicycle), is a device with saddle, pedals, and some form of handlebars arranged as on a bicycle as shown in the figure 3.2. Affixed to exercise bicycle is an ergometer to measure the work done by the exerciser.

Figure 3.3.Recording setup for the physical task stress of the UUM-CycleStress corpus.

3.10 Classifier Settings

For k-Nearest Neighbors (KNN), the value for k used was 2. For Levenberg-Marquardt (LM), learning rate used was 0.1, momentum factor was 0.9 and mean squared error (mse) was set as 0.02% based on 6x10x6 architecture.

CHAPTER FOUR

PARAMETER ANALYSIS

4.0 Introduction

In this chapter, analysis of the extracted speech parameters for both neutral and stressed condition is presented. The analysis is aimed at revealing the effect of physical stress on speech production and its consequent effect on speech acoustic signal. The analysis is made up of heart rate analysis, spectral features of Fundamental frequency (*f0),* Formant frequencies (F1, F2, and F3), spectral Energy and Mel-Frequency Cepstral Coefficients (MFCC).

4.1 Heart rate analysis

Heart rate (HR) readings were taken from 10 speakers made up of 7 males and 3 females at 15 second interval for both the normal and stressed speech data collection process with the aid of Puma heart rate monitor and digital stop watch. Table 4.1 gives summary of speakers' statistics.

Table 4.1

Summary of speakers' statistics

Parameters	Male	Female
No. of speakers	7	3
Speakers age range	24 - 29	27 - 39
Speakers Avg. age	26.5	31.7
No. of words/repetition	5/5	

Native language	Arabic	
Speech style	Isolated word reading	
Sampling rate	8000Hz	
Avg. exertion level	45%	47%

In determining the exertion level of the speakers, the often used formula is that of Karvonen formula (Lemos, Valim, Zandonade, & Natour, 2010; Shnayderman & Katz-Leurer, 2013). Karvonen formula is as given in 4.1

$$HR = (MHR - RHR)\, l + RHR \qquad (4.1)$$

where *HR* is the current heart rate reading taken while exercising, *RHR* is the normal heart rate at resting position, the exertion level l (ranges between 0 to 1) while *MHR* is the speaker's maximum heart rate, which can be calculated by equation 4.2 (Tanaka, Monahan, & Seals, 2001).

$$MHR = 208.9 - 0.7A \qquad (4.2)$$

A denote the age of the speakers.

To apply the formula, average age and average heart rate readings (both normal and stressed conditions) of the speakers based on genders were calculated and applied. From the result as displayed in Table 4.1, average exertion level of 45% and 47% was estimated for male and female respectively. These values represent an appropriate level of exertion. As this low exertion level represent a value closer to naturally valid (Godin et al., 2009).

An interesting fact represented by the exertion Figures, is that exposing males and female to the same physical stress, the female thus exhibit higher exertion level than their male counterpart with 47% as compared to 45% for the males.

Average heart rate was computed for both neutral and stressed speech for each gender. The average heart rate is plotted against time (duration of exercise 10 minutes) as shown in Figure 4.1. It can be observed that while the heart rate level is relatively constant for neutral level, heart rate increases proportionately with the duration of exercise time progression i.e. the longer the exercise time, the higher the heart rate level for the stressed condition. This is similar to result obtained by Godin et al., (2009). However, this work further shows that though both genders when exposed to the same level of stress, females' exertion is higher than that of males' counterpart. This is consistent with the average exertion level determined in Table 4.1 above. Hence, an increase in the level of heart rate during speech is high indication of stress.

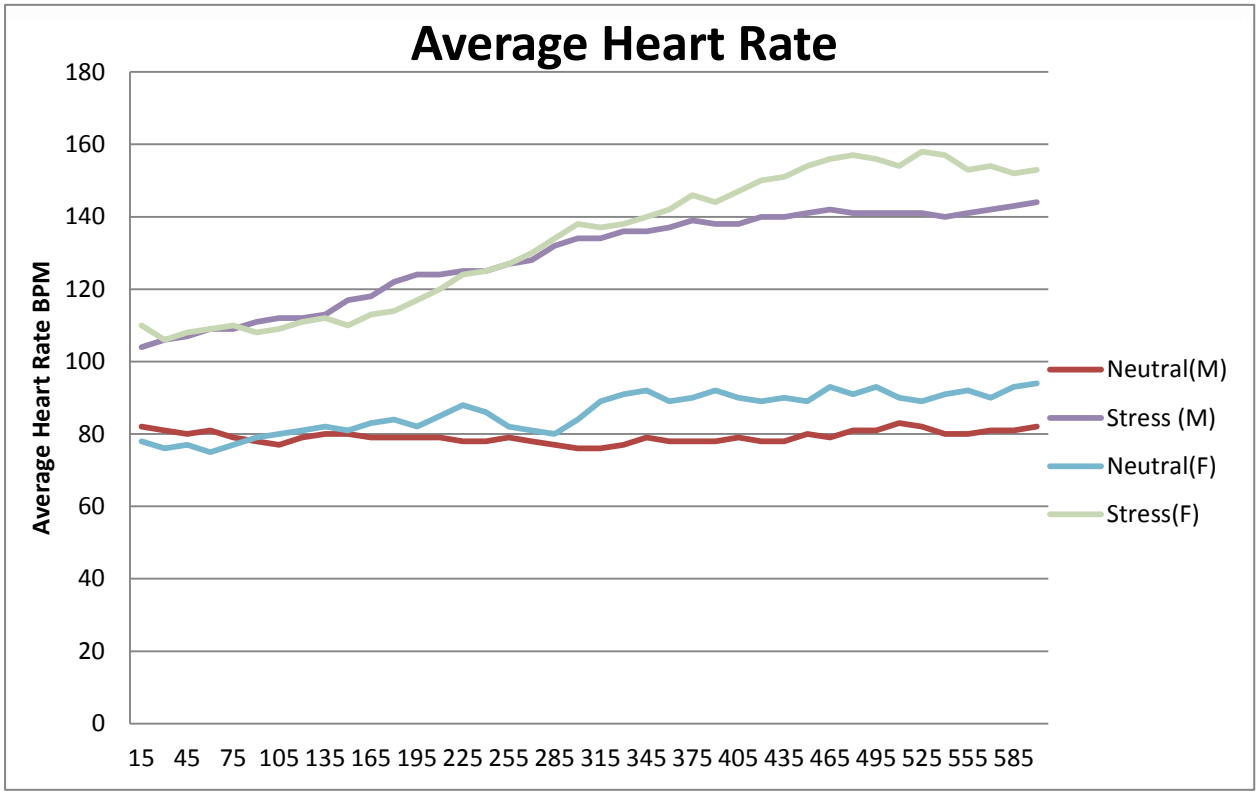


Figure 4.1. Average heart rate for neutral and stress speech for male and female speakers.

4.2 Fundamental frequency (*f0*) Analysis

Fundamental frequency (*f0*) constitute of one the most widely considered spectral features of speech for stress detection (Wang, 2009). Past research shows that the contour of *f0* of an utterance constitute the most sensitive indicator of stress (Hansen et al., 2012). Studies have revealed that changes in *f0* with time, carries ample stress information. It was observed that in a stressful situation, speaker's respiration rate will increase, that will consequently lead to an increase in sub-glottal pressure during utterance. The increase in *f0* during stressed is as a result of increase in sub-glottal pressure. Statistical analysis of *f0* measures such as means, range, max, min, standard deviation and variance of both neutral and stressed speech were computed out as shown in the Table 4.2.

Table 4.2

Statistical measures of fundamental frequency of normal and stressed speech.

Gender	**male**		**Female**	
	Speech type			
Statistical Measures	*Normal*	*Stressed*	*Normal*	*Stressed*
Average	**138.15**	**149.13**	**221.30**	**251.91**
Max	**160**	**181.82**	**250**	**307.69**
Min	**119.403**	**123.08**	**91.95**	**216.22**
Range	**40.60**	**58.74**	**158.05**	**91.477**
Std Dev.	**11.68**	**19.49**	**25.62**	**37.14**
Var.	**136.47**	**379.91**	**656.48**	**1379.196**

Table 4.1 above shows statistical measures of *f0* for both neutral and stressed speech of males and females speakers respectively. The *f0* Figures from the table shows that there is increase

in stressed speech values as compared to neutral speech across all the calculated statistical measures. In conformity with the established theoretical values of *f0*, female's *f0* values are higher than that of the males' counterpart. This is in consistent with (Azmi, 2010). These wide differences between *f0* values of neutral and stressed speech indicate that *f0* is a good indicator of stress speech.

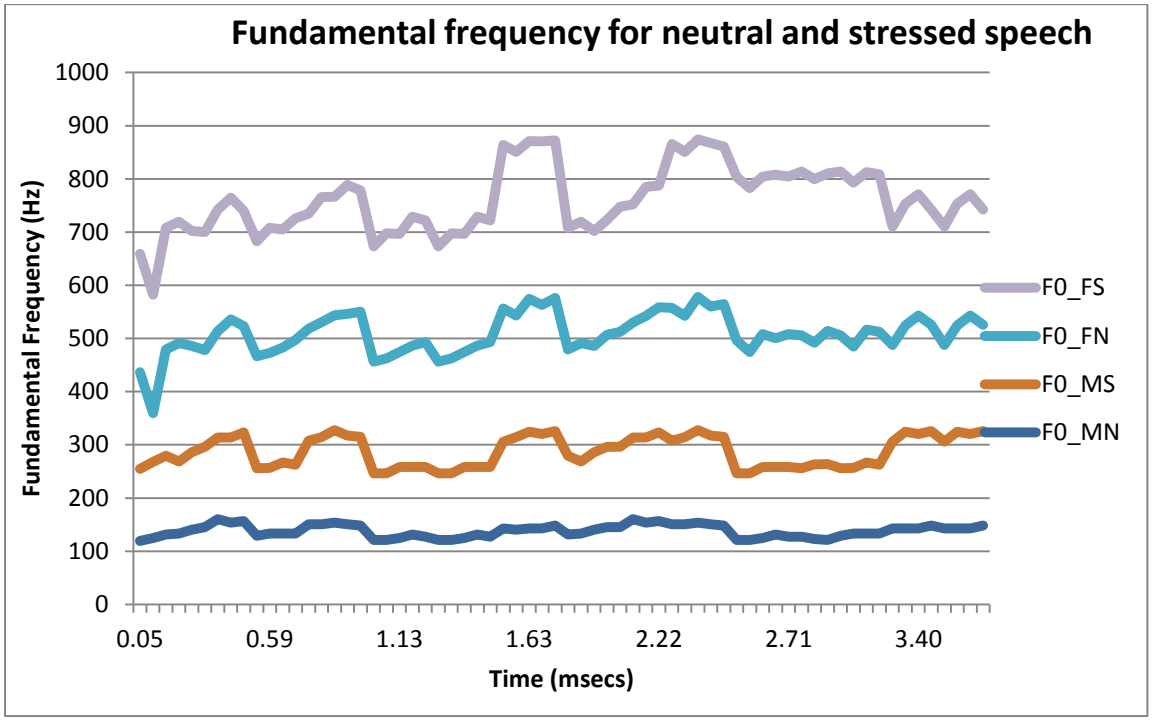


Figure 4.2. Fundamental frequency against time for neutral and stress speech for male and female speakers.

The average values of the calculated *f0* against time for neutral and stress speech for male and female speakers, is plotted as in Figure 4.2 above. The plots reveal clear differences between neutral and stressed speech with stressed speech values higher than neutral speech values across both genders. *f0* values of females are also higher than their male's counterpart in both speech types. From the Figure it can be remark that *f0* asides being a good indicator for stress can also serves to differentiate between genders speech as confirmed in the research of Azmi (2010).

4.3 Formants Analysis

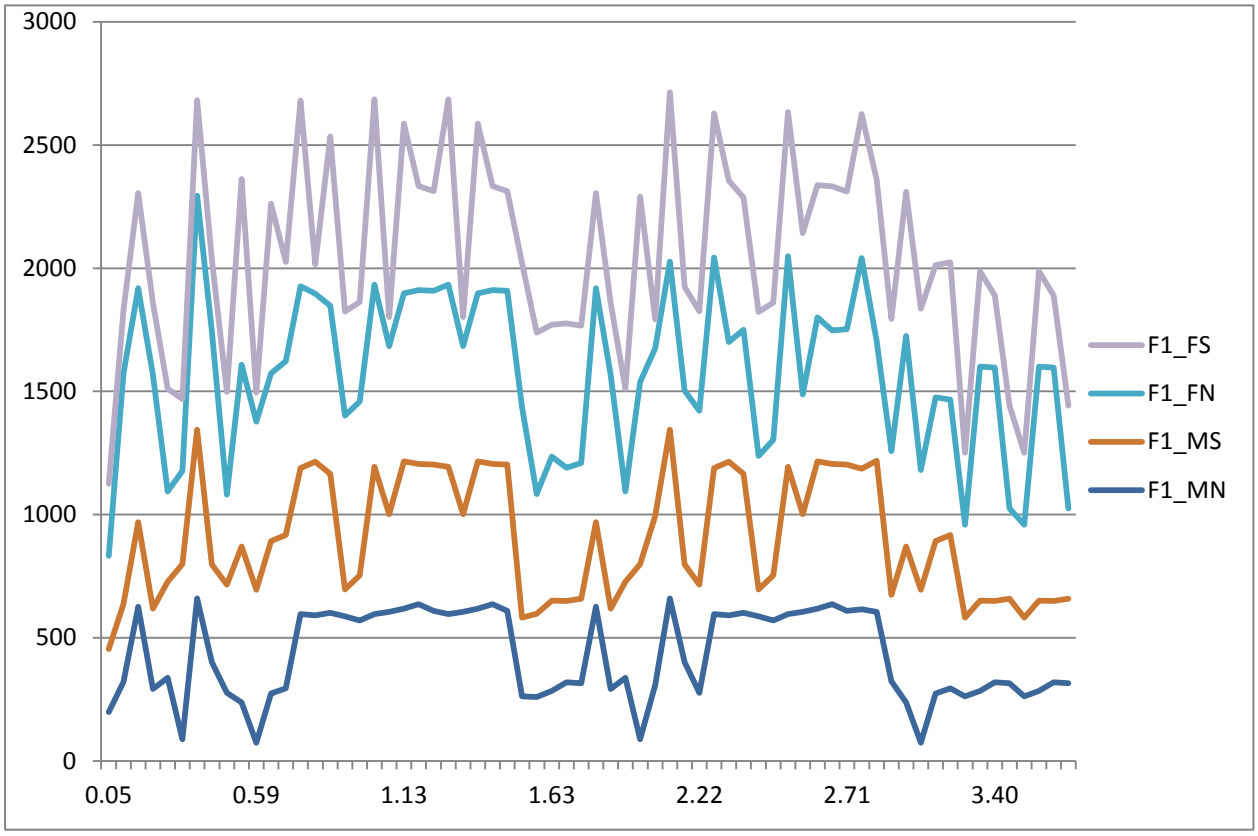
Formant is one of the ways in which stress in speech in manifested. Hence the first three formants F1, F2, and F3 were calculated from neutral and stressed speech. As can be observe from values of F1, F2, and F3 in Tables 4.3 – 4.5, there are changes in the average values of the formants. However these changes are not consistence. While there is increase in the average F1 of stressed speech for male, the female F1 average value shows a decrease Figure for stressed speech. Similar observation can be made for F2 and F3. There is however a trend reversal in average F3 value as stressed value increases for female and shows a decrease for male. As for Figure 4.3 – 4.5, that shows the plots of F1 – F3 for both gender and for neutral and stressed speech. What can be clearly deduced from the Figures. is that the formants for females are consistently higher than that of males for both neutral and stressed speech. However, while in some instances, especially for males, formants values for stressed speech is higher than the neutral speech. The reversal is the case for female formants with stressed speech formants lower than neutral speech formants. Based on these, it can be concluded that formants is not a good indicator for stress. Nonetheless, it can serves to identifier speakers gender as shown by Azmi (2010).

Table 4.3

Statistical measures of first formant (F1) of normal and stressed speech.

Gender	**male**		**Female**	
		Speech type		
Statistical Measures	*Normal*	*Stressed*	*Normal*	*Stressed*
Average	**430.043**	**472.693**	**667.6142**	**473.8286**
Max	**659.2688**	**711.0258**	**949.8152**	**753.1573**
Min	**73.92363**	**109.5532**	**365.4217**	**118.0749**
Range	**585.3451**	**601.4726**	**584.3934**	**635.0824**
Std Dev.	**179.7551**	**156.3661**	**182.9082**	**175.2003**
Var.	**32311.91**	**24450.37**	**33455.41**	**30695.13**

.

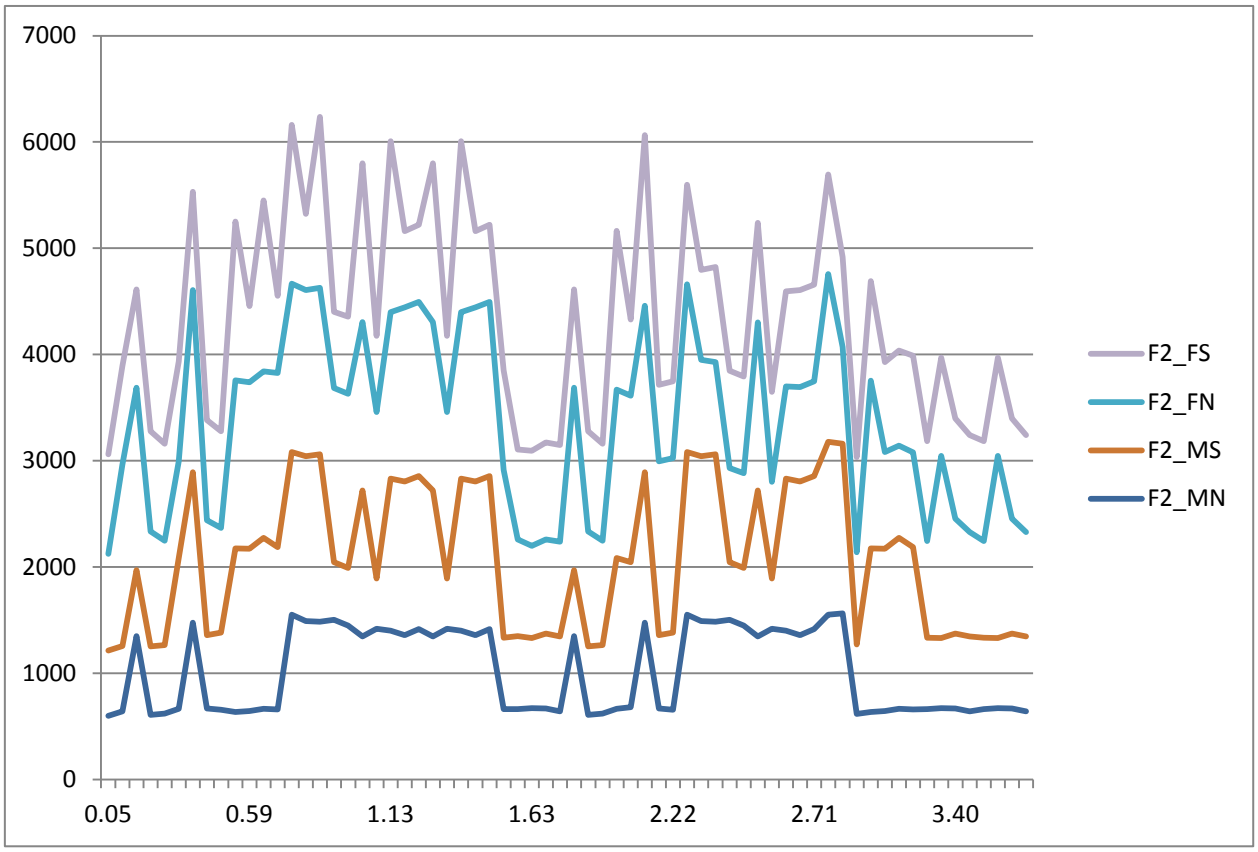


Fig*ure 4.3.* First formant (F1) against time for neutral and stress speech for male and female speaker.

Table 4.4

Statistical measures of second formant (F2) of normal and stressed speech.

Gender	**male**		**Female**	
		Speech type		
Statistical Measures	*Normal*	*Stressed*	*Normal*	*Stressed*
Average	**1021.086**	**1061.487**	**1296.801**	**969.1187**
Max	**1565.082**	**1627.621**	**1715.138**	**1608.573**
Min	**600.3181**	**474.1497**	**867.8964**	**717.0753**
Range	**964.7636**	**1153.471**	**847.2417**	**891.498**
Std Dev.	**398.3421**	**434.4091**	**345.9788**	**261.9818**
Var.	**158676.4**	**188711.3**	**119701.3**	**68634.45**

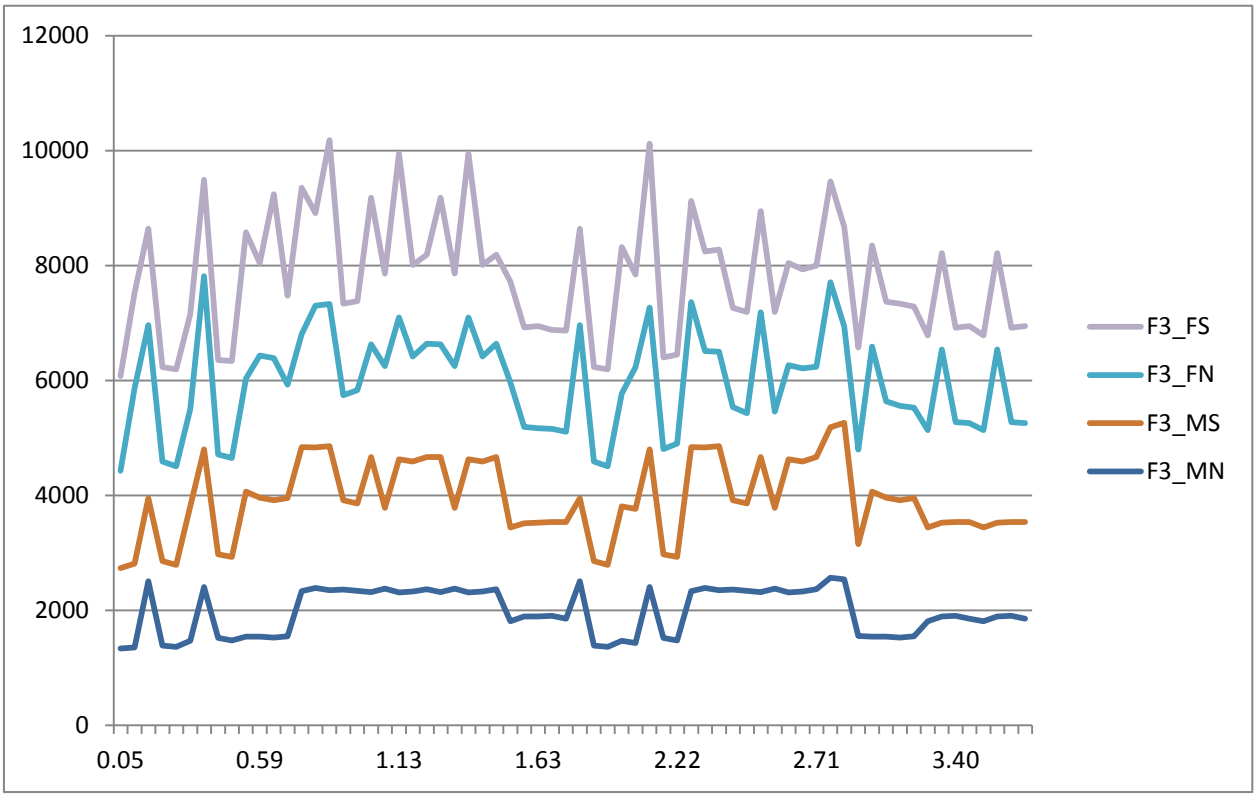


Fig*ure 4.4*. Second formant (F2) against time for neutral and stress speech for male and female speakers.

Table 4.5

Statistical measures of third formant (F3) of normal and stressed speech.

Gender	**male**		**Female**	
			Speech type	
Statistical Measures	*Normal*	*Stressed*	*Normal*	*Stressed*
Average	**1976.356**	**1967.376**	**2033.142**	**1832.451**
Max	**2571.989**	**2724.857**	**3056.764**	**2849.964**
Min	**1338.074**	**1398.674**	**1572.073**	**1546.082**
Range	**1233.916**	**1326.183**	**1484.691**	**1303.882**
Std Dev.	**408.8642**	**447.683**	**452.3847**	**379.107**
Var.	**167169.9**	**200420**	**204652**	**143722.1**



Fig*ure 4.5*. Third formant (F3) against time for neutral and stress speech for male and female speakers.

4.4 Energy

The energy of speech signal over a time period holds enormous information for speech recognition by humans. Hence, given this, it is only reasonable to consider energy of a

frame of speech signal as an additional feature of both neutral and stressed speech. The average, standard deviation and variance energy values of speech are compared across the two types of speech mode.

Table 4.6

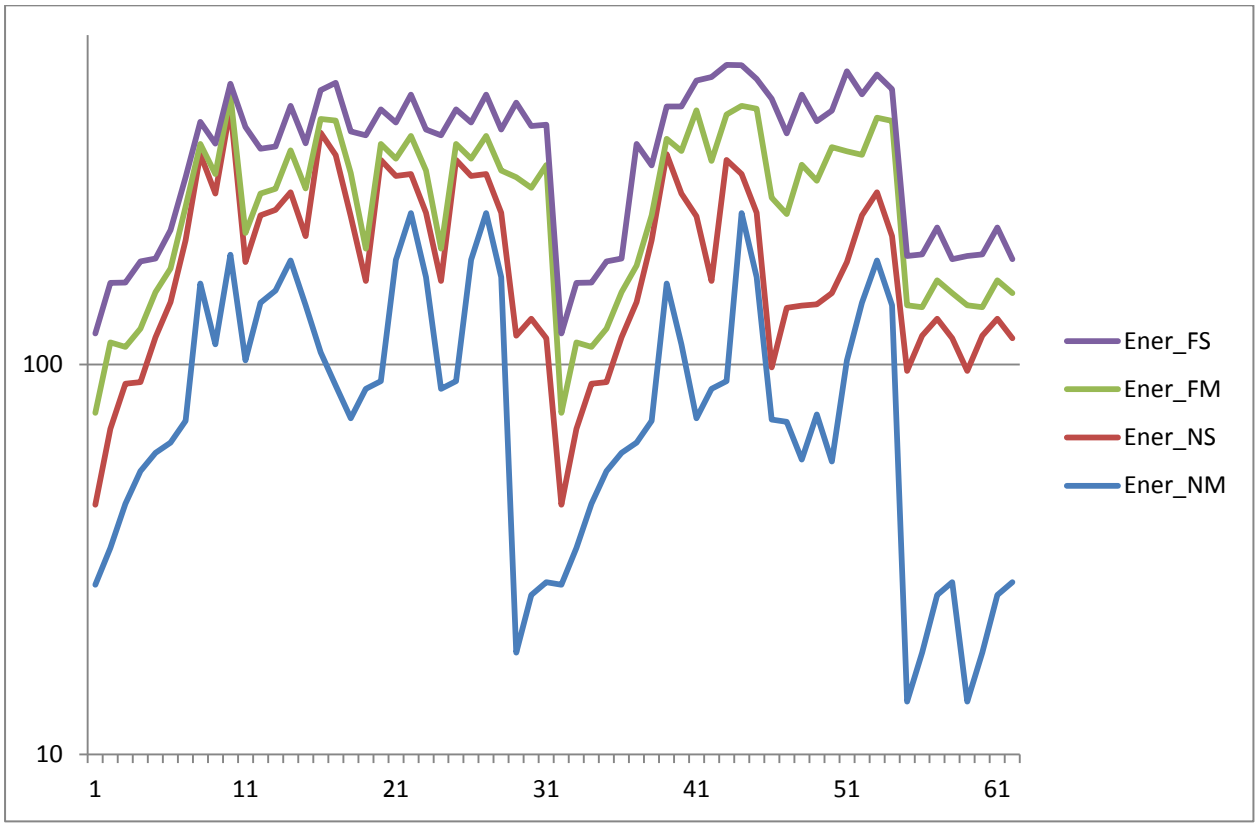
Statistical *measures of energy (dB) of normal and stressed speech.*

Gender	**male**		**Female**	
		Speech type		
Statistical Measures	*Normal*	*Stressed*	*Normal*	*Stressed*
Average	**93.54319**	**103.9178**	**76.13597**	**93.77056**
Max	**244.7369**	**285.5192**	**208.2396**	**212.8721**
Min	**13.67849**	**16.50213**	**21.69867**	**33.6587**
Range	**231.0584**	**269.0171**	**186.5409**	**179.2135**
Std Dev.	**63.09371**	**62.80181**	**61.22617**	**50.43013**
Var.	**3980.817**	**3944.068**	**3748.644**	**2543.198**

From Table 4.6 above, there are clear differences in the average values of both neutral and stressed speech across both genders. The average value of stressed speech is higher than that of neutral speech. Indicating that there is correlation between stressed and increase in energy level of speech. The higher the change in energy level, the higher the level of stressed for the particular speaker. However for both standard deviation and variance, there is decrease in the observed values of stressed speech as compare to the neutral speech across both genders.

In furtherance to these, Figure. 4.6 below shows the relationship between speech energy level over time for neutral and stressed speech signals. As expected, the value of stressed energy is higher for both genders with female values higher than the male values. From both the Table 4.6 and Figure 4.6, it can be concluded that thought there is difference in energy

values between neutral and stressed speech signal, however the differences are not significant enough.



Fig*ure 4.6.* Energy against time for neutral and stress speech for male and female speakers.

4.5 Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) is the most commonly used features for speech recognition. In this work the first 13- coefficients of MFCC is extracted from speech signal while NN was used in classification. The NN was trained using Levenberg-Marquardt back propagation (LM). The classification result results so obtained from the training were compared with those obtained from other classifiers such as Linear Discriminant Analysis (LDA) and K-Nearest Neighbor (KNN). Several combinations of coefficients and parameters were varied. Classification results are group into two major classes based on gender and words recognition. The classification results are presented in the following sections.

4.5.1 Classification result by Gender

The first set of classification experiments were carried out based on the 13-MFCC and 4-MFCC coefficients. The obtained results were as presented.

Table 4.7 and Figure 4.7 above show the CR% of 13-MFCC coefficient for neutral and stressed speech for male and female speakers. Based on the Classification Rate (CR %) Figures, ANN have the lowest Classification Rate (CR%) followed by LDA while KNN performs better. ANN attains an average of 24.15% and 21.46% for neutral and stressed speech respectively for males. Similarly 21.64% recognition rate was attained by ANN for both neutral and stressed speech of females. For KNN, 81.35% and 40.76% was achieved for neutral and stressed speech of male while 70.49% and 35.33% was achieved for female speech. As for LDA 71.10% and 36.36% CR for male and; 65.65% and 32% for female for neutral and stressed speeches respectively. From the Classification Rate (CR %) values across the three classifiers, it can be observed that neutral speech has a higher recognition rate than that stressed speech. Also, male speech was recognized well than the female speech.

Table 4.7

Classification result of 13-MFCC for normal and stressed speech.

	13-MFCC coefficients Classifiers					
Speech type/Gender	*ANN*		*K-NN*		*LDA*	
	Male	*Female*	*Male*	*Female*	*Male*	*Female*
Neutral	**24.15**	**21.64**	**81.35**	**70.49**	**71.10**	**65.65**
Stressed	**21.46**	**21.64**	**40.76**	**35.33**	**36.36**	**32.00**
Overall	**22.08**	**21.64**	**61.06**	**52.91**	**53.73**	**48.82**

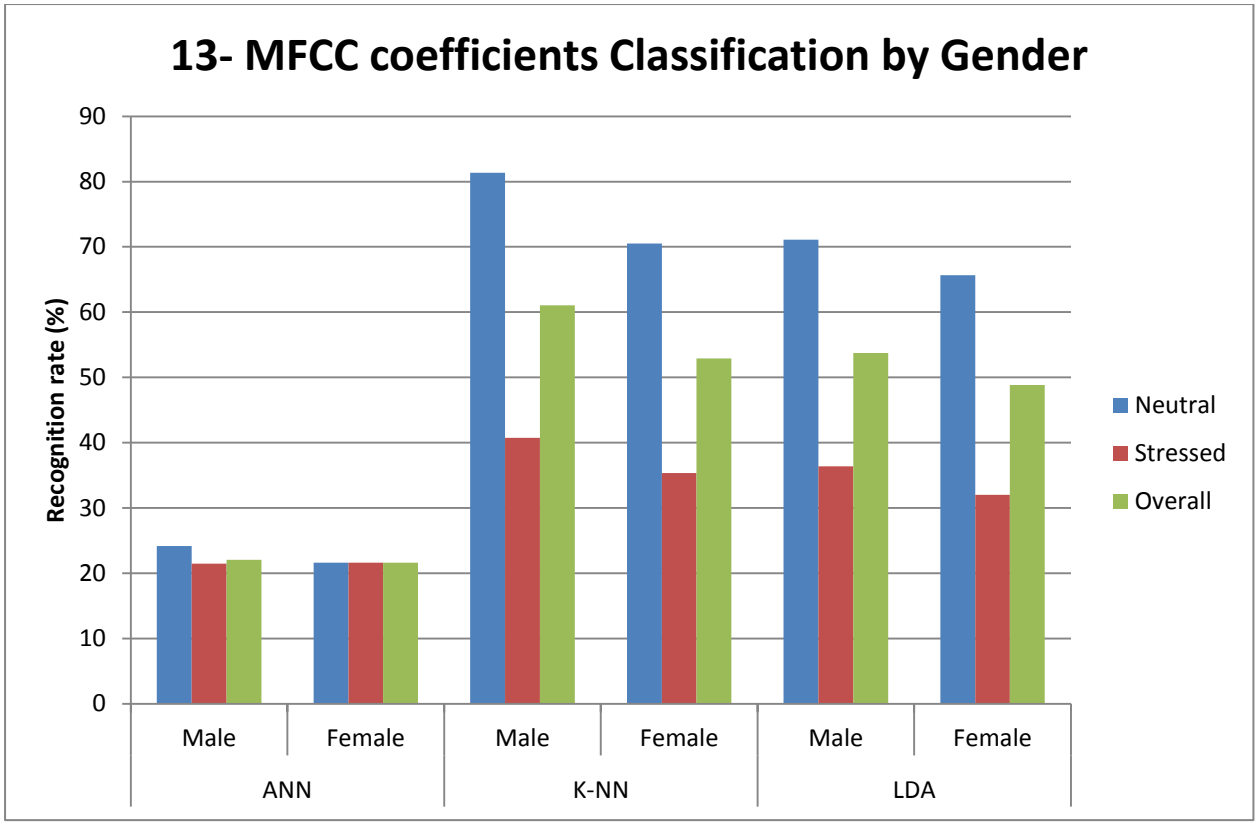


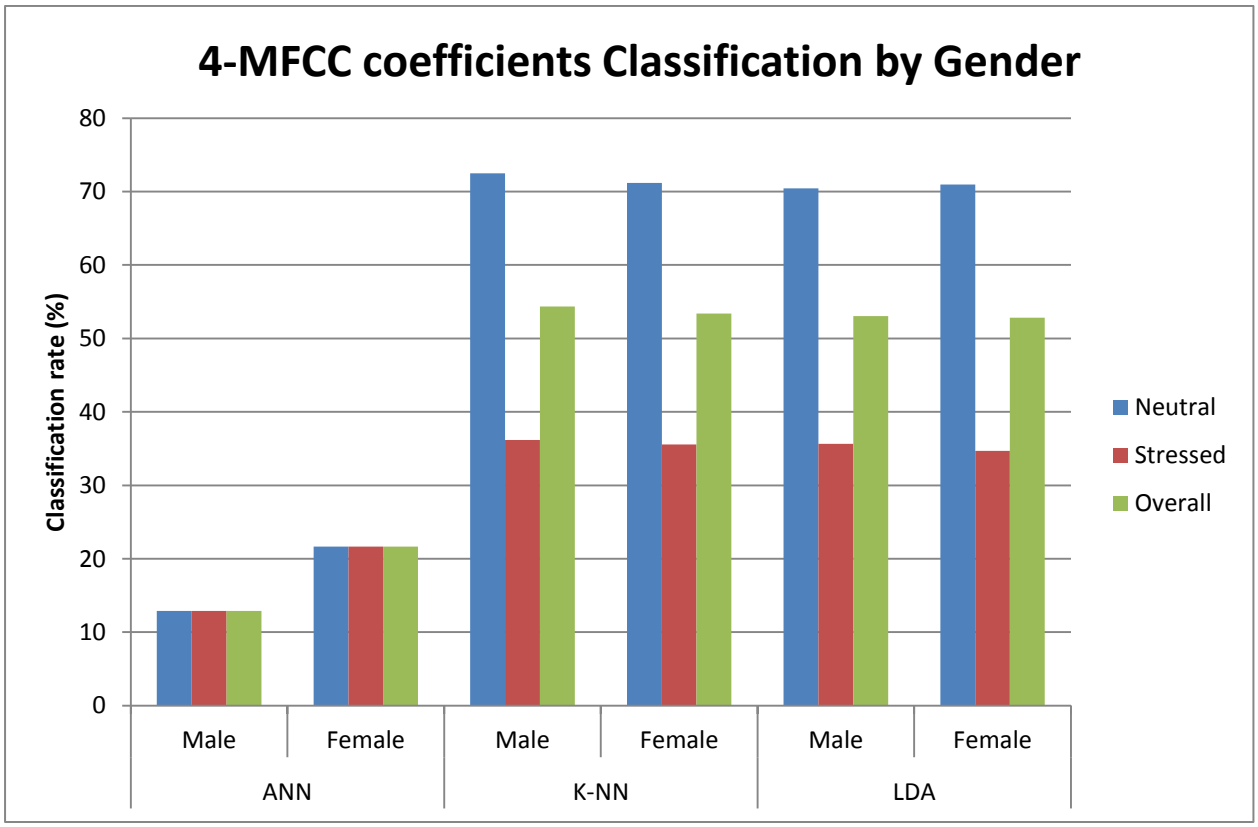
Figure 4.7. Classification Rate (CR %) chart of 13-MFCC for neutral and stress speech for male and female speakers.

Shown in Table 4.8 and Figure 4.8, the Classification Rate (CR %) values of 4-MFCC for neutral and stressed speech of both male and female. Generally, 4-MFCC Classification Rate CR% is lower than that of 13-MFCC coefficients. Similar to the results obtained for 13MFCC, ANN has the lowest recognition rate of 12.9% and 21.64 for both neutral and speech stressed for male and female respectively. KNN Classification Rate (CR%) is 72.5% and 71.20% for neutral for both genders respectively. 36.19% and 35.57% for stressed speech of both genders. LDA has 70.42% and 70.96% for neutral speech of both genders. 35.66% and 34.67% for stressed speech of both genders.

Table 4.8

Classification result of 4-MFCC for normal and stressed speech.

	4-MFCC coefficients Classifiers					
Speech type/Gender	*ANN*		*K-NN*		*LDA*	
	Male	Female	Male	Female	Male	Female
Neutral	**12.90**	**21.64**	**72.50**	**71.20**	**70.42**	**70.96**
Stressed	**12.90**	**21.64**	**36.19**	**35.57**	**35.66**	**34.67**
Overall	**12.90**	**21.64**	**54.35**	**53.39**	**53.04**	**52.82**



Fig*ure 4.8.* Classification Rate (CR %) chart of 4-MFCC for neutral and stress speech for male and female speakers.

4.5.2 Classification result by words

Further recognition experiments were performed based on word levels with 13 and 4 MFCC coefficients. The word "Bow" was dropped from the experiment after the MFCC coefficients for all the female speakers shows the same value.

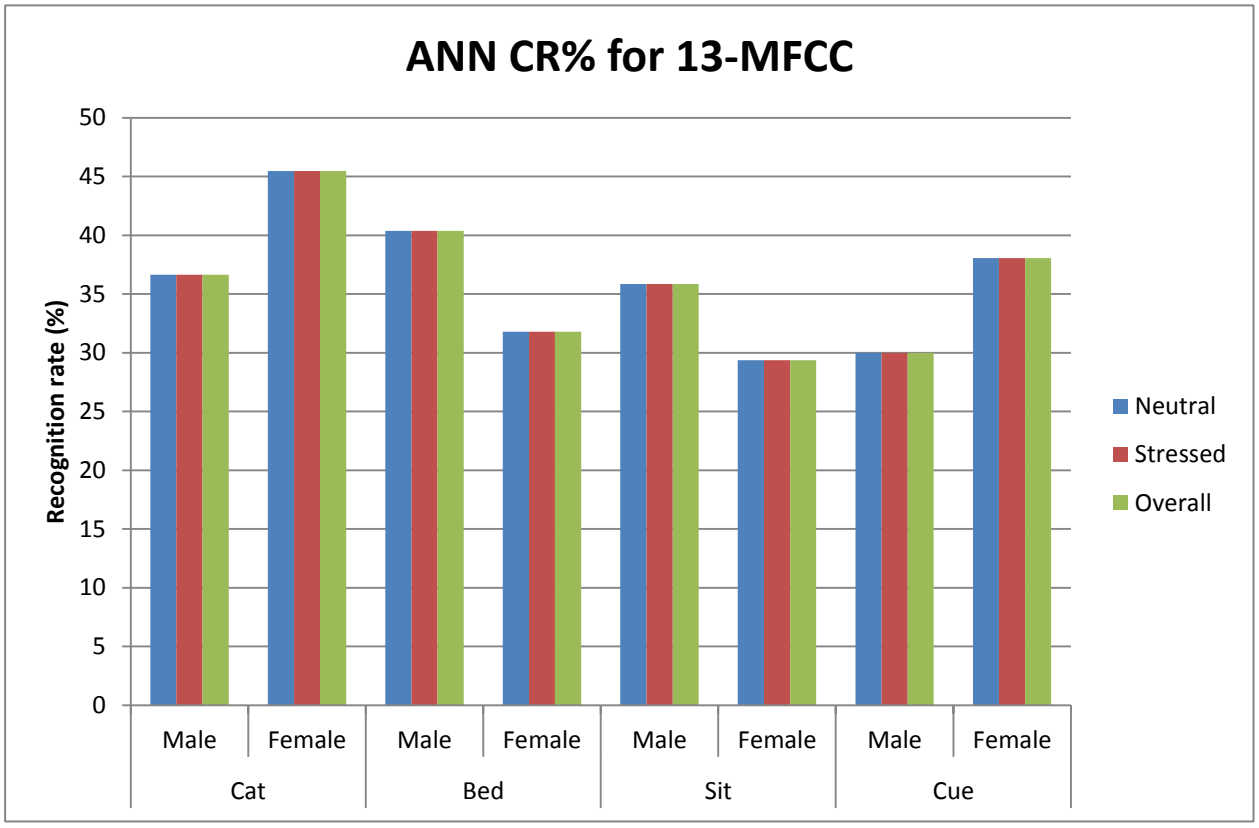
4.5.2.1 Classification result by words with 13 MFCC

Classification results of ANN, LDA, and KNN for both gender and speech type were as displayed in the tables that follows. Tables 4.9 – 4.11 and Figure 4.9 – 4.11 show the average values and plots of 13-MFCC classification results for ANN, KNN, and LDA respectively. As in the previous results ANN has the lowest word recognition rate follow by LDA, while KNN has the highest Classification Rate (CR %) value. For ANN, Classification Rate (CR%) ranges between 29 – 45%, with "cat" being the most recognized with 45.48% and "sit" the lowest with 29.37%. As for KNN, Classification Rate CR% values range between 30 – 94%. "Cue" has the highest Classification Rate CR% of 66% and "bed" the lowest with 55%. LDA Classification Rate CR% values ranges 34 – 84% with "cue" having the highest Classification Rate (CR%) of 61% and "Sit" the lowest classification rate with 57% recognition rate.

Table 4.9

Word based ANN classification result of 13-MFCC normal and stressed speech.

	Cat		**Bed**		**Sit**		**Cue**	
Gender	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*
Neutral	**36.64**	**45.48**	**40.38**	**31.79**	**35.86**	**29.37**	**30**	**38.06**
Stressed	**36.64**	**45.48**	**40.38**	**31.79**	**35.86**	**29.37**	**30**	**38.06**
Overall	**36.64**	**45.48**	**40.38**	**31.79**	**35.86**	**29.37**	**30**	**38.06**



Fig*ure 4.9.* ANN Word based Classification Rate (CR %) chart of 13-MFCC for neutral and stress speech.

Table 4.10

Word based KNN classification result of 13-MFCC normal and stressed speech.

	Cat		**Bed**		**Sit**		**Cue**	
Gender	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*
Neutral	**79.67**	**89.59**	**79.84**	**67.02**	**82.95**	**94.94**	**85.98**	**93.68**
Stressed	**40.34**	**44.71**	**40.33**	**30.80**	**37.67**	**44.20**	**42.86**	**42.71**
Overall	**60.01**	**67.15**	**60.09**	**48.91**	**60.31**	**69.57**	**64.42**	**68.20**

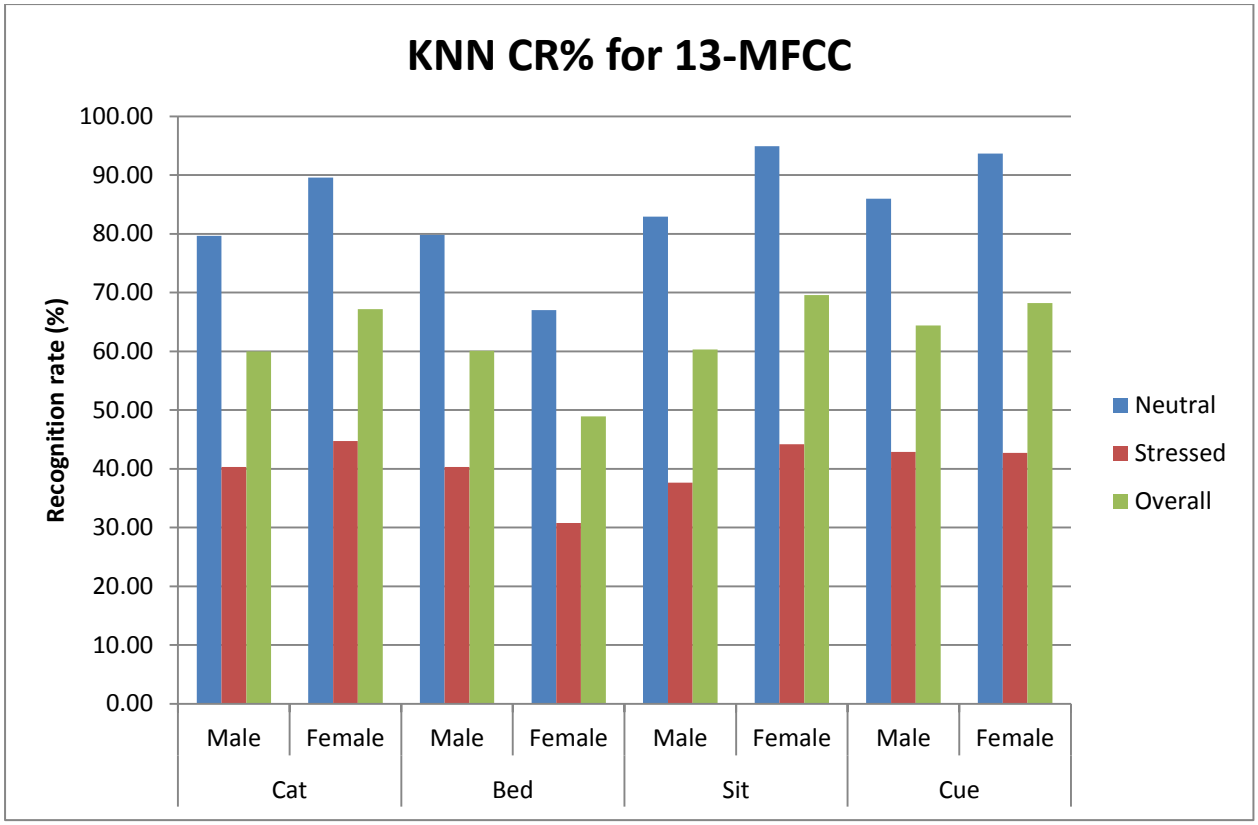
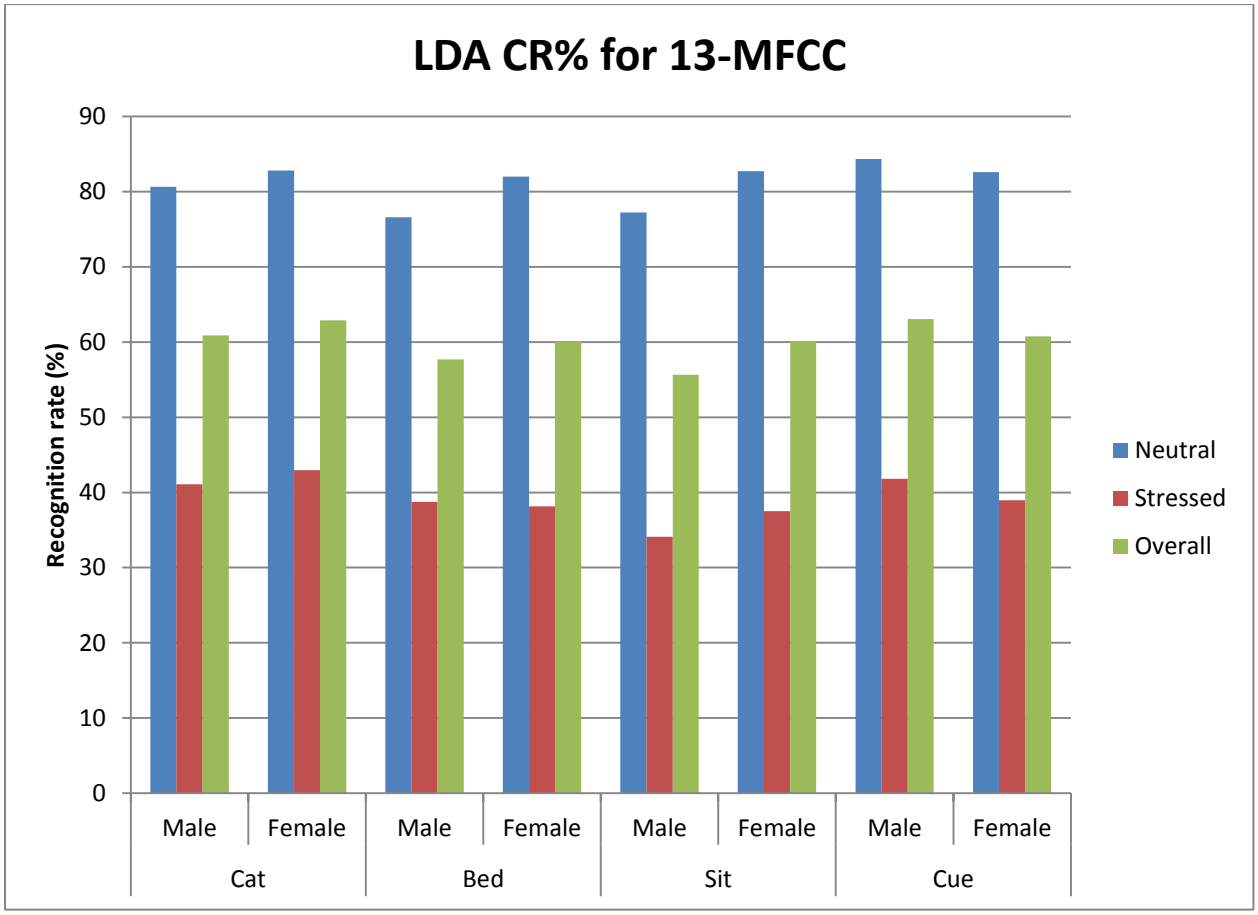


Figure 4.10. KNN Word based Classification Rate (CR %) chart of 13-MFCC for neutral and stress speech.

Table 4.11

Word based LDA classification result of 13-MFCC normal and stressed speech.

	Cat		**Bed**		**Sit**		**Cue**	
Gender	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*
Neutral	**80.61**	**82.8**	**76.58**	**82**	**77.21**	**82.71**	**84.32**	**82.56**
Stressed	**41.1**	**42.97**	**38.75**	**38.13**	**34.09**	**37.5**	**41.79**	**38.94**
Overall	**60.855**	**62.885**	**57.665**	**60.065**	**55.65**	**60.105**	**63.055**	**60.75**



Fig*ure 4.11*. LDA Word based Classification Rate (CR %) chart of 13-MFCC for neutral and stress speech.

4.5.2.2 Classification result by words with 4 MFCC

Classification results of ANN, LDA, and KNN with 4-MFCC for both gender and speech type were as displayed in the tables that follows. Tables 4.12 – 4.14 and Figures 4.12 – 4.14 show the average values and plots of 4-MFCC classification results for ANN, KNN, and LDA respectively. As in the previous results ANN has the lowest word recognition rate follow by LDA, while KNN has the highest Classification Rate CR% value. For ANN, Classification Rate (CR%) ranges between 32.64% for "cat" spoken by male in stressed mode to 47.67% for "cat" spoken by female in neutral mode. Both the neutral and stressed speech of a gender has the same Classification Rate CR% value for a given word. Highest recognition value is for the word "cat" spoken by female in neutral mode with 47.67%. Based on this result, it can

be generalized that females Classification Rate CR% in higher than males Classification Rate CR%.

In the case of KNN classifier, Classification Rate CR% values ranges between 31.82% for word "bed" spoken by male in stressed mode, 82.56% for "sit" spoken by male in neutral mode. The highest recognition rate of 65.31% is for word "sit" while "bed" has the lowest recognition rate of 55.05%. Males have higher Classification Rate CR% value for words: cat, cue, and sit while females got the highest Classification Rate CR% for word bed.

As for LDA, Classification Rate (CR%) values ranges from 27.75% for stressed "bed" as spoken by male to 80.74% for "cue" spoken by female in neutral mode. Females Classification Rate CR% has higher values than male's Classification Rate CR%. Likewise, neutral speech has higher Classification Rate CR% values than stressed speech. "Sit" was recognized with highest Classification Rate CR% of 64% while "bed" has the lowest Classification Rate (CR %) value of 48.26%. The consistent higher Classification Rate CR% value of females can be employed in designing adaptive ASR based of LDA classifier.

Table 4.12

Word based ANN classification result of 4-MFCC normal and stressed speech.

	Cat		**Bed**		**Sit**		**Cue**	
Gender	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*
Neutral	**32.64**	**47.67**	**22.75**	**46.11**	**42.12**	**46.57**	**37.86**	**43.1**
Stressed	**32.64**	**47.67**	**22.75**	**46.11**	**42.12**	**46.57**	**37.86**	**43.1**
Overall	**32.64**	**47.67**	**22.75**	**46.11**	**42.12**	**46.57**	**37.86**	**43.1**

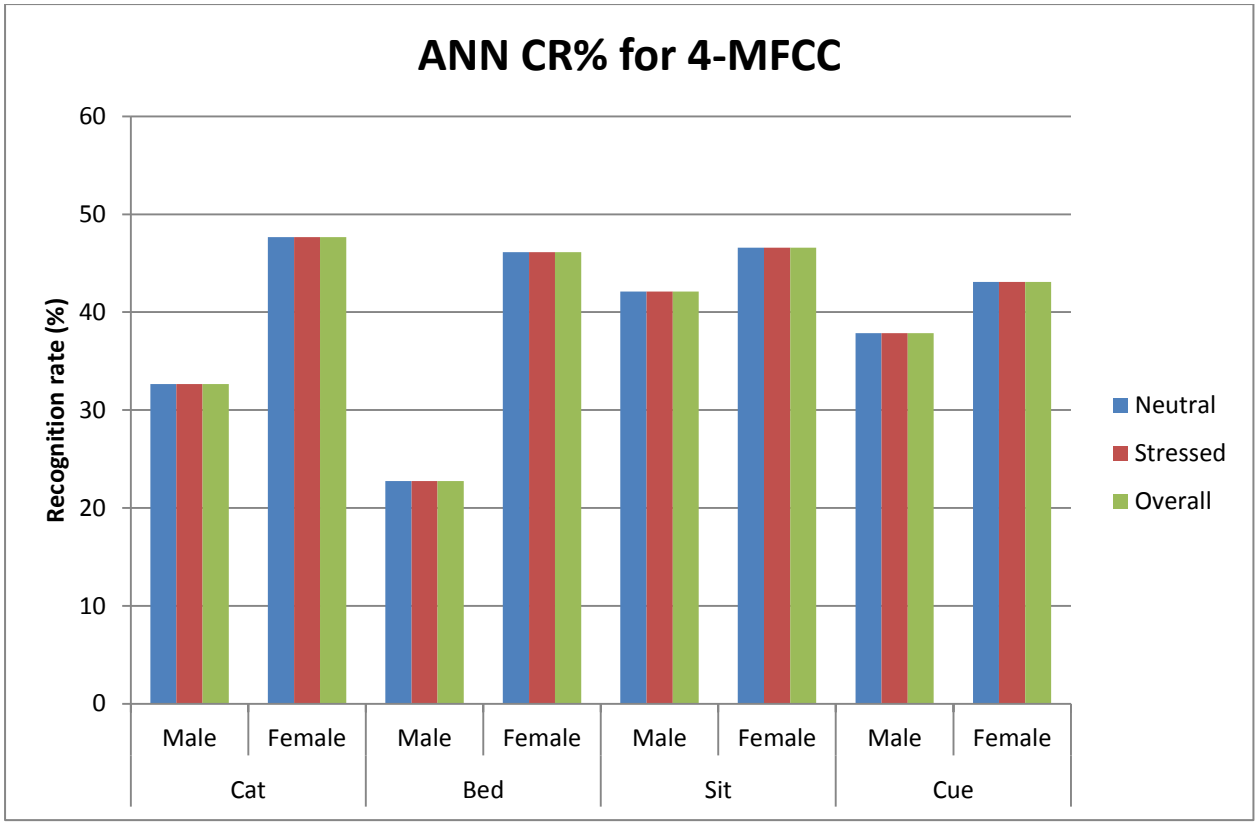


Figure 4.12. ANN Word based Classification Rate (CR %) chart of 4-MFCC for neutral and stress speech.

Table 4.13

Word based KNN classification result of 4-MFCC normal and stressed speech.

	Cat		**Bed**		**Sit**		**Cue**	
Gender	Male	Female	Male	Female	Male	Female	Male	Female
Neutral	**77.03**	**76.87**	**69**	**77.9**	**82.56**	**80.72**	**77.42**	**74.93**
Stressed	**39.04**	**38.5**	**31.82**	**41.48**	**61**	**36.96**	**38.62**	**40.22**
Overall	**58.035**	**57.685**	**50.41**	**59.69**	**71.78**	**58.84**	**58.02**	**57.575**

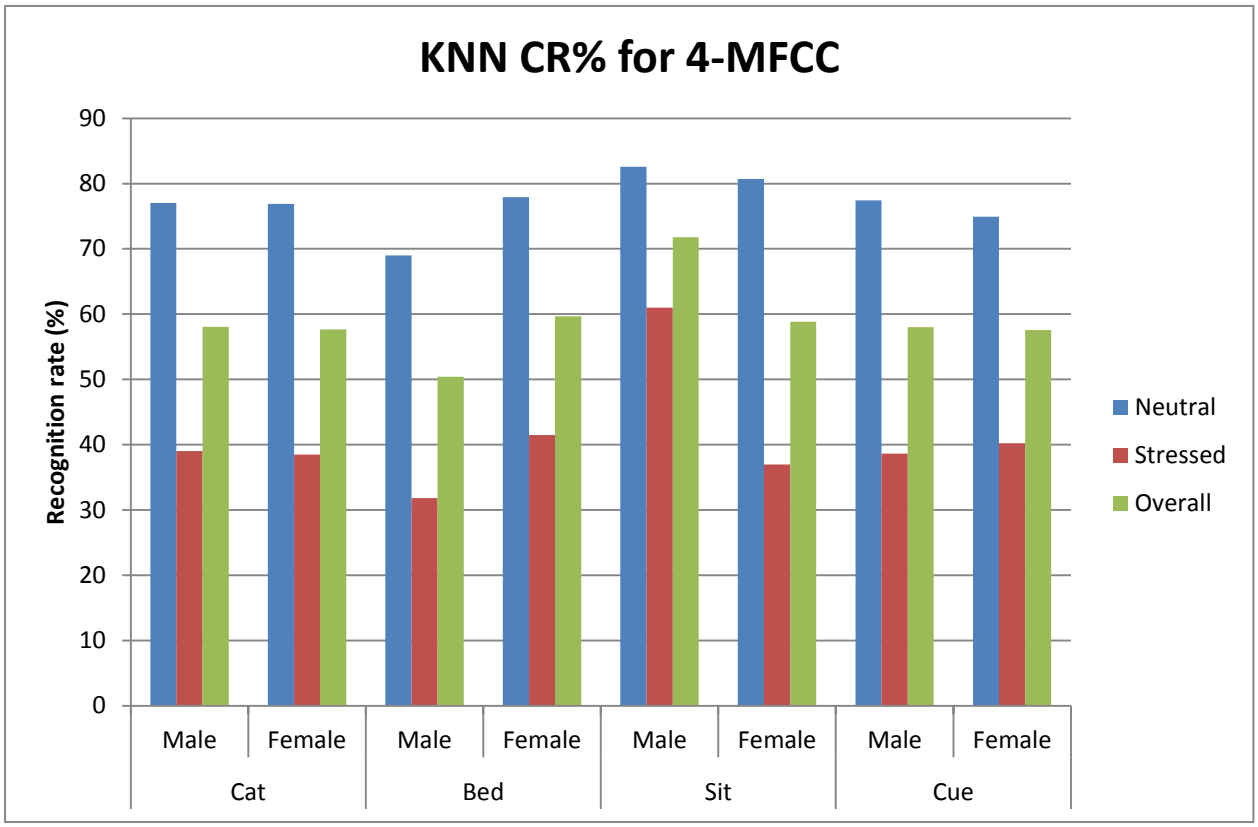
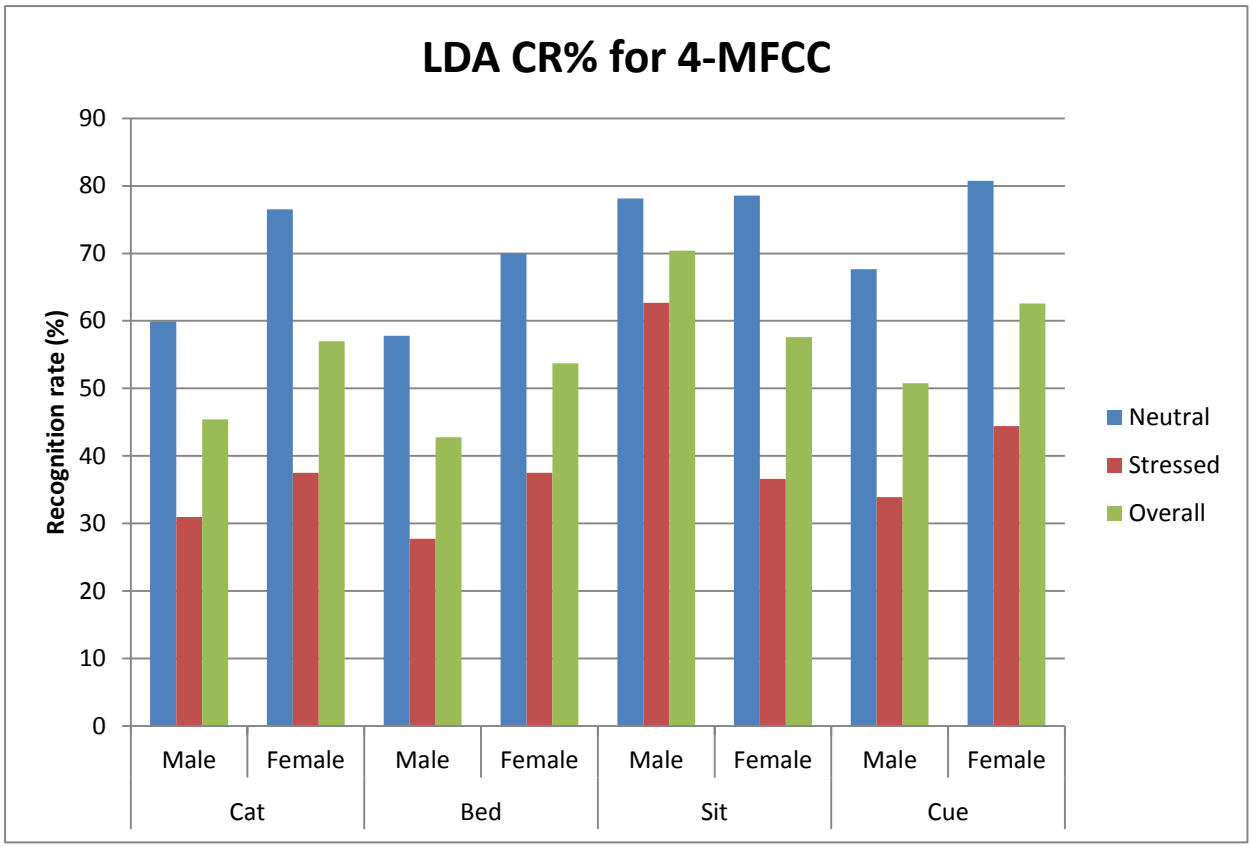


Figure 4.13. KNN Word based Classification Rate (CR %) chart of 4-MFCC for neutral and stress speech.

Table 4.14

Word based LDA classification result of 4-MFCC normal and stressed speech.

	Cat		**Bed**		**Sit**		**Cue**	
Gender	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*
Neutral	**59.93**	**76.52**	**57.81**	**69.96**	**78.11**	**78.57**	**67.64**	**80.74**
Stressed	**30.94**	**37.5**	**27.75**	**37.5**	**62.67**	**36.61**	**33.91**	**44.44**
Overall	**45.435**	**57.01**	**42.78**	**53.73**	**70.39**	**57.59**	**50.775**	**62.59**



Fig*ure 4.14*. LDA Word based CR% chat of 4-MFCC for neutral and stress speech.

4.6 Stress classification

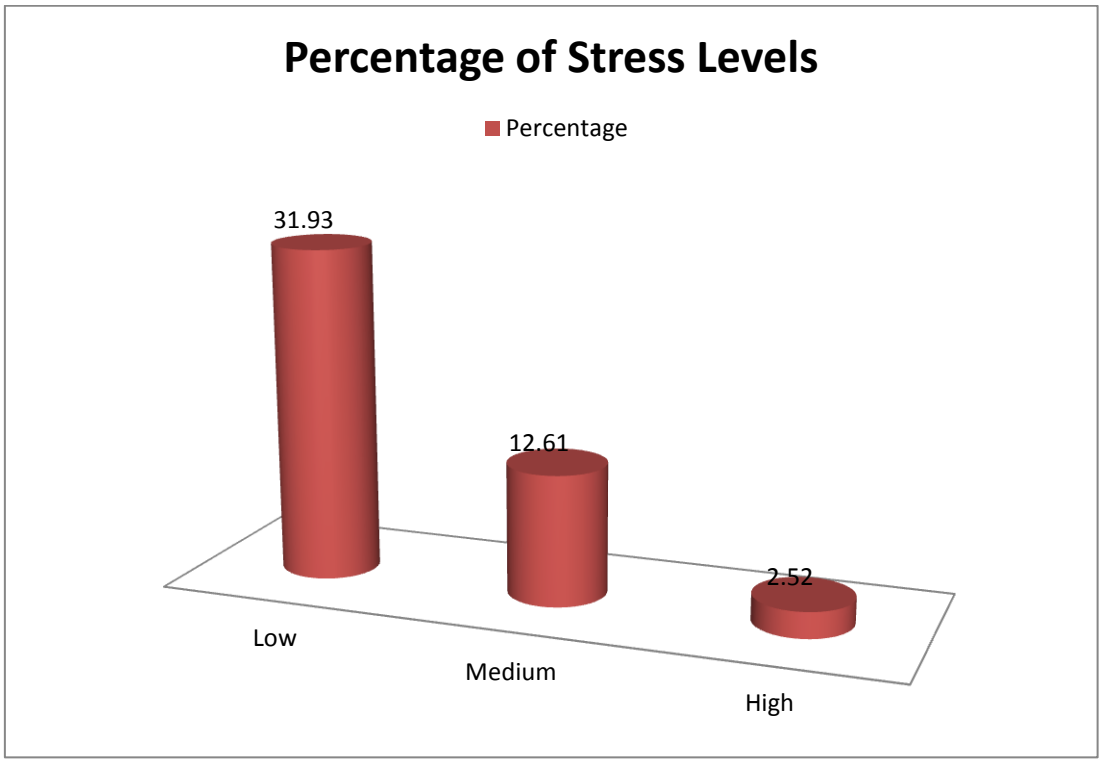
Furtherance to the objectives of this thesis is the classification of stressed levels. After successful recognition of speech into neutral or stressed, then comes the classification of stressed speech into different levels of: low, medium, and high. Stress classification process is based on the outcome of classification result of KNN since it gives a higher classification output. Based on the output of the classification result of KNN, data corresponding to stressed speech is examined and thresholds are formulated for stress level classification. The result of stress level classification is as shown in the Table 4.15 and Figure 4.15 below. Table 4.15 shows the three levels of stress and their corresponding percentage based on the number of stressed instances recognized by KNN classifier. From the table, 31.93%, 12.61%, and 2.52% are classified as low, medium and high stress level respectively. This classification of

stress level is vividly displayed in the bar chart in Figure 4.15. The classification rate of stress level is slow may be due to the amount of sample data used in this study is low.

Table 4.15

Stress Level and percentage of stressed speech.

Stress Level	Percentage
Low	**31.93**
Medium	**12.61**
High	**2.52**



Fig*ure 4.15.* Percentage classification of stress level of stressed speech.

CHAPTER FIVE

CONCLUSION AND FUTURE RESEARCH

5.0 Introduction

This chapter marks the conclusion of the research work. It contains the discussions of results and the conclusions, and the contributions of the study. Recommendations further research in the future to improve this research work was also given.

5.1 Conclusion

Going by the objectives of this research work, which includes the identification of variant features of stressed speech, determination of speech features most affected by stress, and to classify stress levels in stressed speech. This work achieved all the objectives which discussion follows.

Heart rate analysis was conducted to determine the effect of physical exercise on speech production. The result of the analysis reveals that physical exercise led to an increase in the heart rate levels and which consequently affect speech production. The heart rate level increases proportionately with the duration of exercise time progression i.e. the longer the exercise time, the higher the heart rate level for the stressed condition. Results also show that exposing males and females to the same level of physical exercise resulted into different exertion effects with female exertion level higher than the male with 47% for female as against 45% for males.

Parametric analysis of speech features of *f0,* F1, F2, F3, and energy was performed in determining the effect of stress on these features. Based on statistical average results

conducted on these features, the results show that *f0* is highly affected by stress. The average of *f0* shows significant increases from neutral to stressed speech. The average *f0* values of 138 and 149Hz for neutral and stressed speech for males while 221and 251Hz for neutral and stressed speech for females respectively were obtained from experimental. This result is consistence with the theory that female has a higher *f0* than the male and also consistence with the findings of (Godin et al., 2009). As for the first three formants of F1, F2, and F3, the experimental analysis shows that there were changes in the values of the formants between neutral and stressed speech. The changes are however not consistence across the gender and the formants. The changes alternate between positive and negative for the different formant and across the gender. Hence, though stress causes changes in formants levels, it is however insufficient for the purpose of classification of stress levels. As for speech energy level, the results show that the average value of stressed speech is higher than that of neutral speech. Indicating that there is correlation between stressed and increase in energy level of speech. The higher the change in energy level, the higher the level of stressed for the particular speaker. As expected, the value of stressed energy is higher for both genders with female values higher than the male values. It can be concluded that thought there is difference in energy values between neutral and stressed speech signal, however the differences are not significant enough. Based on the findings above, it can be stated that increase in heart rate, energy, *f0*, and the first three formants F1, F2, and F3 can be directly attributed to physical exercise undertook by the subjects. This implies that physical exercise can induced stress in individual. Hence, the first objective of this thesis is achieved. It can also be deduced that of all the features examined, *f0* is the most correlated with stressed. Therefore it fulfils the second objective of being the most affected feature by stress.

Two variants of MFCC coefficients 13-MFCC and 4-MFCC were used in classification. Three sets of classifiers: ANN, KNN, and LDA. Results of classifications were based on cross validation techniques whereby the database is randomly divided into training and testing sets based on the ratio of 7:3. A total of 30-fold cross validations tests were done and their averaged classification results were computed for each of the classifier.

The results for 13-MFCC features for all the words in the corpus based on CR% shows that ANN has the lowest CR% due to the fact that it requires a large volume of data for both training and classification. LDA followed with a medium CR% while KNN has the highest CR% because of its ability to perform well with minimal data. This high recognition rate of KNN is comparable to what is reported by Yusof and Yacoob (2008). ANN attains an average of 24.15% and 21.46% for neutral and stressed speech respectively for males. Similarly 21.64% recognition rate was attained by ANN for both neutral and stressed speech of females. For KNN, 81.35% and 40.76% was achieved for neutral and stressed speech of male while 70.49% and 35.33% was achieved for female speech. As for LDA 71.10% and 36.36% CR for male and; 65.56% and 32% for female for neutral and stressed speeches respectively. From the CR% values across the three classifiers, it can be observed that neutral speech has a higher recognition rate than that stressed speech. Also, male speech was recognized well than the female speech.

While for 4-MFCC features CR% is lower than that of 13-MFCC coefficients. Similar to the results obtained for 13MFCC, ANN has the lowest recognition rate of 12.9% and 21.64 % for both neutral and speech stressed for male and female respectively. KNN CR% is 72.5% and 71.20% for neutral for both genders respectively. 36.19% and 35.57% for stressed speech

of both genders. LDA gave 70.42% and 70.96% for neutral speech of both genders. 35.66% and 34.67% for stressed speech of both genders.

For word recognition experiments with 13-MFCC classification results, ANN has the lowest word recognition rate follow by LDA, while KNN has the highest CR% value. For ANN, CR% ranges between 29 – 45% with "cat" being the most recognized with 45.48% and "sit" the lowest with 29.37%. As for KNN, CR% values range between 30 – 94%. "Cue" has the highest CR% of 66% and "bed" the lowest with 55%. LDA CR% values ranges 34 – 84% with "cue" having the highest CR% of 61% and "Sit" with 57% recognition rate. Likewise for 4-MFCC word based classification, for ANN, both the neutral and stressed speech of a gender has the same CR% value for a given word. Highest recognition value is for the word "cat" spoken by male in neutral mode with 47.67%. Based on this result, it can be generalized that females CR% in higher than males CR%. In the case of KNN classifier, the highest recognition rate of 65.31% is for word "sit" while "bed" has the lowest recognition rate of 55.05%. Males have higher CR% value for words: cat, cue, and sit while females got the highest CR% for word bed. As for LDA, Females CR% has a higher values than males CR%. Likewise, neutral speech has higher CR% values than stressed speech. "Sit" was recognized with highest CR% of 64% while "bed" has the lowest CR% value of 48.26%. The consistent higher CR% value of females can be employed in designing adaptive ASR based of LDA classifier. Aside the recognition of speech into neutral and stressed, further classification of stressed speech into three categories of low, medium, and high was carried out. Based on the result, 31.93% of recognized stressed speech is classified as low, 12.61% as medium, and 2.52% as high stress respectively. With this the third and final objective of classifying stress into different levels is achieved. In conclusion, this research work has achieved all the set objectives.

5.3 Future research

It is evident from the results of this research work that physical exercise induces stress and which consequently resulted into variation in speech production. This research also demonstrated the ability of using speech processing technique to detect and classify stress levels in humans. However, to further enhance the applicability of ASR for stress detection, further studies is required. These includes but not limited to:

a) Do establish the causes of inconsistency in the values of the first three formants and energy across gender.
b) To experiment with other speech extraction techniques such as LPC and other classifiers like Support Machine Vector (SVM) and many other Neural Network algorithm.
c) To experiment with other languages and also a mixtures of languages.

REFERENCES

Abu Shariah, M., Ainon, R. N., Zainuddin, R., & Khalifa, O. O. (2007). *Human computer interaction using isolated-words speech recognition technology.*Paper presented at the Intelligent and Advanced Systems, 2007. ICIAS 2007. International Conference on.

Acero, A.(1990). *Acoustical and environmental robustness in automatic speech recognition* (Doctoral dissertation, Carnegie Mellon University).

Adami, A. G., Lazzarotto, G. B., Foppa, E. F., & Couto Barone, D. (1999). *A comparison between features for a residential security prototype based on speaker identification with a model of artificial neural network.* Paper presented at the Computational Intelligence and Multimedia Applications, 1999. ICCIMA'99. Proceedings. Third International Conference on.

Amuda, S., Boril, H., Sangwan, A., & Hansen, J. H. (2010). *Limited resource speech recognition for Nigerian English.* Paper presented at the Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.

Anusuya, M., & Katti, S. (2010). Speech recognition by machine, A review. *arXiv preprint arXiv:1001.2267*.

Anusuya, M., & Katti, S. (2011). Classification Techniques used in Speech Recognition Applications: A Review. *Int. J. Comp. Tech. Appl, 2*(4), 910-954.

Azmi, M. M., & Tolba, H. (2008). *Noise robustness using different acoustic units.* Paper presented at the Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on.

Azmi, M. Y. S., Idayu, M. N., Roshidi, D., Yaakob, A., & Yaacob, S. (2012). Noise Robustness of Spectrum Delta (SpD) Features in Malay Vowel Recognition. In *Computer Applications for Communication, Networking, and Digital Contents* (pp. 270-277): Springer.

Bakker, J., Holenderski, L., Kocielnik, R., Pechenizkiy, M., & Sidorova, N. (2012). *Stess@ work: From measuring stress to its understanding, prediction and handling with personalized coaching.* Paper presented at the Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., et al. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication, 49*, 763-786.

Bou-Ghazale, S. E., & Hansen, J. H. (2000). A comparative study of traditional and newly proposed features for recognition of speech under stress. *Speech and Audio Processing, IEEE Transactions on, 8*(4), 429-442.

Buchanan, C. R. (2005). Informatics Research Proposal-Modelling the Semantics of Sound. *School of Informatics, University of Edinburgh, United Kingdom*.

Casale, S., Russo, A., Scebba, G., & Serrano, S. (2008). *Speech emotion classification using machine learning algorithms.* Paper presented at the Semantic Computing, 2008 IEEE International Conference on.

Chee, L. S., Ai, O. C., Hariharan, M., & Yaacob, S. (2009, November). MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA. In *Research and Development (SCOReD), 2009 IEEE Student Conference on* (pp. 146-149). IEEE.

Cohen, S. E., Kessler, R. C., & Gordon, L. U. E. (1995). *Measuring stress: A guide for health and social scientists*: Oxford University Press.

Costello, A., Abbas, M., Allen, A., Ball, S., Bell, S., Bellamy, R., et al. (2009). Managing the health effects of climate change: lancet and University College London Institute for Global Health Commission. *The Lancet, 373*(9676), 1693-1733.

Daniel, J., & Martin, J. (2009). Speech and Language Processing-Edition: 2. *Prentice-Hall Inc, ISBN, 131873210*, 2-16,489.

Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *Audio,Speech And Language Processing, IEEE Transactions on; 21*(5), 1-30.

Dhole, N. P., & Gurjar, A. A. (2013). Detection of Speech under Stress: A Review. *International Journal of Engineering and Innovative Technology (IJEIT) on* (Vol. 2, issue 10, pp. 36-38).

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. *New York: John Wiley, Section, 10*, l.

Entwistle, M. S., & Adviser-Granaas, M. (2005). *Training methods and enrollment techniques to improve the performance of automated speech recognition systems under conditions of human exertion*: University of South Dakota.

Godin, K. W., Hansen, J. H., Busso, C., & Katz, W. F. (2009). Classification based analysis of speech under physical task stress. *Master's thesis, University of Texas at Dallas, Richardson, TX*.

Gray, S. S. (2006). *Speech Science Modeling for Automatic Accent and Dialect Classification.* University of Colorado.

Han, Z., Lung, S., & Wang, J. (2012, March). A study on speech emotion recognition based on ccbc and neural network. In *Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on* (Vol. 2, pp. 144-147). IEEE.

Hansen, J. H., Sangwan, A., & Kim, W. (2012). Speech under stress and Lombard effect: impact and solutions for forensic speaker recognition. In *Forensic Speaker Recognition* (pp. 103-123): Springer.

Haykin, S. S. (2009). *Neural networks and learning machines* (Vol. 3): Prentice Hall New York.

He, L., Lech, M., Maddage, M. C., & Allen, N. (2009, August). Stress detection using speech spectrograms and sigma-pi neuron units. In *Natural Computation, 2009. ICNC'09. Fifth International Conference on* (Vol. 2, pp. 260-264). IEEE.

Hong, J.-H., Ramos, J., & Dey, A. K. (2012). *Understanding physiological responses to stressors during physical activity.* Paper presented at the Proceedings of the 2012 ACM Conference on Ubiquitous Computing.

Huang, X., Acero, A., & Hon, H. W. (2001). *Spoken language processing* (Vol. 18). Englewood Cliffs: Prentice Hall.

Ibiyemi, T., & Akintola, A. (2012). Automatic Speech Recognition for Telephone Voice Dialling in Yorùbá. *International Journal of Engineering, 1*.

Khalifa, O. O., El-Darymli, K. K., Abdullah, A.-H., & Daoud, J. I. (2013). Statistical Modeling for Speech Recognition.

Khanna, P., & Kumar, M. S. (2011). Application of Vector Quantization in Emotion Recognition from Human Speech. In *Information Intelligence, Systems, Technology and Management* (pp. 118-125). Springer Berlin Heidelberg.

Koolagudi, S. G., Kumar, N., & Rao, K. S. (2011, February). Speech emotion recognition using segmental level prosodic analysis. In *Devices and Communications (ICDeCom), 2011 International Conference on* (pp. 1-5). IEEE.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech using source, system, and prosodic features. *International Journal of Speech Technology*, *15*(2), 265-289.

Kurniawan, H., Maslov, A. V., & Pechenizkiy, M. (2013). *Stress detection from speech and Galvanic Skin Response signals.* Paper presented at the Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on.

Lai, M., Chen, Y., Chu, M., Zhao, Y., & Hu, F. (2006). *A hierarchical approach to automatic stress detection in English sentences.* Paper presented at the Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on.

Lemos, M. C. D., Valim, V., Zandonade, E., & Natour, J. (2010). Intensity level for exercise training in fibromyalgia by using mathematical models. *BMC musculoskeletal disorders, 11*(1), 54.

Levit, M., Huber, R., Batliner, A., & Noeth, E. (2001). *Use of prosodic speech characteristics for automated detection of alcohol intoxication.* Paper presented at the ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding.

Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., et al. (2012). *StressSense: Detecting stress in unconstrained acoustic environments using smartphones.* Paper presented at the Proceedings of the 2012 ACM Conference on Ubiquitous Computing.

Mao, X., Chen, L., & Fu, L. (2009, March). Multi-level speech emotion recognition based on HMM and ANN. In *Computer Science and Information Engineering, 2009 WRI World Congress on* (Vol. 7, pp. 225-229). IEEE.

Martin, R. (2005). Statistical methods for the enhancement of noisy speech. In *Speech Enhancement* (pp. 43-65): Springer.

Mohd Yusof, S., & Yaacob, S. (2008). Classification of Malaysian vowels using formant based features. *Journal of ICT, 7*, 27-40.

Murray, I. R., Baber, C., & South, A. (1996). Towards a definition and working model of stress and its effects on speech. *Speech Communication, 20*(1), 3-12.

Nakatani, T., Juang, B.-H., Kinoshita, K., & Miyoshi, M. (2005). *Harmonicity based dereverberation with maximum a posteriori estimation.* Paper presented at the Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on.

Narayana, M., & Kopparapu, S. (2009). *On the use of stress information in speech for speaker recognition.* Paper presented at the TENCON 2009-2009 IEEE Region 10 Conference.

Neely, S. T., & Allen, J. B. (1979). Invertibility of a room impulse response. *The Journal of the Acoustical Society of America, 66*, 165.

Patil, S. A., & Hansen, J. H. (2008). Detection of speech under physical stress: Model development, sensor selection, and feature fusion.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals*(Vol. 100, p. 17). Englewood Cliffs: Prentice-hall.

Rao, K. S., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology*, *16*(2), 143-160.

Razak, Z., Ibrahim, N. J., & Idna Idris, M. (2008). Quranic Verse recitation recognition module for support in J-QAF learning: A Review. *International Journal of Computer Science and Network Security (IJCSNS), 8*(8), 207-216.

Scherer, S., Hofmann, H., Lampmann, M., Pfeil, M., Rhinow, S., Schwenker, F., et al. (2008). *Emotion Recognition from Speech: Stress Experiment.* Paper presented at the LREC.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., & Wendemuth, A. (2009). *Acoustic emotion recognition: A benchmark comparison of performances.* Paper presented at the Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on.

Shnayderman, I., & Katz-Leurer, M. (2013). An aerobic walking programme versus muscle strengthening programme for chronic low back pain: a randomized controlled trial. *Clinical rehabilitation*, *27*(3), 207-214.

Sigmund, M. (2010). *Changes in frequency spectrum of vowels due to psychological stress.* Paper presented at the Radioelektronika (RADIOELEKTRONIKA), 2010 20th International Conference.

Sigmund, M., & Dostal, T. (2004). Analysis of emotional stress in speech. *Proc. IASTED AIA 2004*, 317-322.

Siraj, F., Shahrul Azmi, M., Paulraj, M., & Yaacob, S. (2009). *Malaysian Vowel Recognition Based on Spectral Envelope Using Bandwidth Approach.* Paper presented at the Modelling & Simulation, 2009. AMS'09. Third Asia International Conference on.

Sjölander, K., & Beskow, J. (2000). *Wavesurfer-an open source speech tool.* Paper presented at the INTERSPEECH.

Sun, Z., Yuan, X., Bebis, G., & Louis, S. J. (2002). *Neural-network-based gender classification using genetic search for eigen-feature selection.* Paper presented at the Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on.

Tanaka, H., Monahan, K. D., & Seals, D. R. (2001). Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology*, *37*(1), 153-156.

Torabi, S., AlmasGanj, F., & Mohammadian, A. (2008, December). Semi-Supervised Classification of Speaker's Psychological Stress. In *Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International* (pp. 1-4). IEEE.

Wang, Y. (2009). *Speech recognition under stress.* Southern Illinois University Carbondale.

Wu, M., & Wang, D. (2006). A two-stage algorithm for one-microphone reverberant speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on, 14*, 774-784.

Yan, Q., Vaseghi, S., Rentzos, D., & Ho, C.-H. (2007). Analysis and synthesis of formant spaces of British, Australian, and American accents. *Audio, Speech, and Language Processing, IEEE Transactions on, 15*, 676-689.

Yegnanarayana, B., & Murthy, P. S. (2000). Enhancement of reverberant speech using LP residual signal. *Speech and Audio Processing, IEEE Transactions on, 8*, 267-281.

You, H., & Adviser-Alwan, A. (2009). *Robust automatic speech recognition algorithms for dealing with noise and accent*: University of California at Los Angeles.

Zhai, J., & Barreto, A. (2008). Stress detection in computer users through non-invasive monitoring of physiological signals. *Blood, 5*, 0.

Zhai, J., Barreto, A. B., Chin, C., & Li, C. (2005). *Realization of stress detection using psychophysiological signals for improvement of human-computer interactions.* Paper presented at the SoutheastCon, 2005. Proceedings. IEEE.

Zhang, H. (2012). Emotional Speech Recognition Based on Syllable Distribution Feature Extraction. In *Foundations of Intelligent Systems* (pp. 415-420). Springer Berlin Heidelberg.

Zhou, G., Hansen, J. H., & Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *Speech and Audio Processing, IEEE Transactions on, 9*(3), 201-216.