EARLY PREDICTION OF CLINICAL DEPRESSION IN ADOLESCENTS USING SINGLE-CHANNEL AND MULTI-CHANNEL CLASSIFICATION APPROACH

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

Ooi Kuan Ee, Brian

B.Eng. Electronics

School of Electrical and Computer Engineering
Science, Engineering and Technology Portfolio

RMIT University

March 2014
ABSTRACT

Given the systematically growing prevalence of clinical depression in adolescents and the burden on economies worldwide caused by this debilitating state, automated, easy to use early prediction techniques are sought after for efficient mass screening and prevention programs. Existing preventive measures are restricted to conventional psychological assessment based on questionnaires and interviews.

The thesis examined the possibility of using fully automated speech or image analysis to predict onset of Major Depression in adolescents 2.5 years before the symptoms meet the conventional diagnostic criteria.

The prediction task was facilitated by the nature of available audiovisual recordings made when all participants were professionally diagnosed as normal healthy adolescents with no current or previous episodes of depression. These recordings were accompanied by clinical diagnosis made 2.5 years after the recordings were taken; this diagnosis showed which participants developed symptoms of Major Depression. Using this information, it was possible to classify the audiovisual recordings into two categories: participants “At Risk” (AR) for depression and participants “Not At Risk” (NAR) for depression. The study was based on 15 participants (6 male and 9 female) representing the AR group and 15 participants (6 male and 9 female) representing the NAR group. To our knowledge, this was the first study of its kind.

The experimental results led to the following major conclusions:

1) Facial images are less effective than acoustic speech characteristics in prediction of depression.
2) Acoustic speech analysis can provided efficient prediction of risk for depression 2.5 years before the full blown symptoms occur. The experiments showed that the speech based classification accuracy can reach 74%.

3) The efficiency of individual types of features was tested using a classical single-channel classification approach. It was found that out of the four categories (glottal, prosodic, TEO and spectral), the glottal features were the most efficient in discriminating between AR and NAR individuals. The best prediction accuracy provided by the glottal features was 69% with good sensitivity to specificity ratio of 76%/62%.

4) The study proposed a new multi-channel weighted speech classification (MCWSC) method. It was found that the optimised version of this method, called the optimised multi-channel weighted speech classification (OMCWSC) system was the most efficient in prediction of risk for depression. When implemented in the person based (PB), two weights (2W) approach with 3 single-channels processing glottal (G), prosodic (PS) and TEOG (within 1300Hz-5500Hz) features, the OMCWSC achieved 74% prediction accuracy, 77% sensitivity and 70% specificity.
DECLARATION

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

Ooi Kuan Ee, Brian

March 2014
ACKNOWLEDGEMENTS

The work presented here would have not been possible without the endless help and support from a few people which has been part of this journey together with me. It has been a long and though period but the overwhelming support have played a big part in carrying me through these times and have heavily motivated this adventure of my life. Thus, I would like to express my deepest appreciation to all these people.

To my main supervisor Dr. Margaret Lech, you have been the most supportive person throughout this research project. First of all, I would like to thank you for providing me with this opportunity and believing in me to work together with you on this project. Your expert opinions and ideas that you have provided me are very much appreciated and valued. Thank you for your tireless guidance and patience that you have shown me in every area of my study.

I also like to thank Professor Nicholas Allen from the School of Psychological Science of the University of Melbourne, for the time and advice he has provided especially in relation to the psychology side of this research. Thank you for your effort in providing me with the helpful information and ideas that I can work on the database from our collaborative work. Also special thanks to Dr. Lisa Sheeber and her team for the database provided to us.

Not to forget, Dr. Alex Low from the Wolfson Microelectronics, who has been a great friend and colleague in our DSP lab. It has been a great adventure together and I would like to thank you for kindness in helping me settle in to the research environment and providing me with valuable advice from your experience. I truly value the time we
spend together as work mates that made my time working in the lab so much more exciting.

I am also very thankful for the special experience that I share with my partner who went through the same path of pursuing her doctoral studies at the same time as me. It is a blessing that we can share and understand our experience that we go through and encourage each other.

Finally, I would like to thank my family for their unconditional love and support. Thank you for the constant calls to check on my progress and see how I am going. Your constant prayers and encouragement has kept me going and not giving up. Thanks mom and dad for always believing in me.
LIST OF PUBLICATIONS

Book Chapters:


Journal Publications:


Conference Publications:


TABLE OF CONTENTS

ABSTRACT ........................................................................................................................ ii
DECLARATION.................................................................................................................... iv
ACKNOWLEDGEMENTS .................................................................................................. v
LIST OF PUBLICATIONS .............................................................................................. vii
TABLE OF CONTENTS ................................................................................................. viii
LIST OF TABLES ............................................................................................................. xi
LIST OF FIGURES AND ILLUSTRATIONS ................................................................ xiii

CHAPTER ONE: .................................................................................................................1
1.1 Preview ......................................................................................................................1
1.2 Background and Problem Statement ....................................................................1
1.3 Thesis Aims ..............................................................................................................5
1.4 Thesis Scope and Limitations .................................................................................5
1.5 Thesis Contributions ...............................................................................................7
1.6 Thesis Structure .......................................................................................................8

CHAPTER TWO: ..............................................................................................................10
2.1 Preview ..................................................................................................................10
2.2 Clinical Depression .............................................................................................11
2.3 Depression in Adolescents ..................................................................................12
2.4 Detection Versus Prediction .................................................................................14
2.5 Depression Recognition from Facial Images .......................................................16
   2.5.1 Representation of Facial Expressions .........................................................17
   2.5.2 Facial Image Analysis and Classification ..................................................18
2.6 Depression Recognition from Speech ..................................................................22
   2.6.1 Physiological basis for speech based diagnosis .........................................22
   2.6.2 Speech acoustics as correlates of Depression ............................................23
   2.6.3 Studies of speech based diagnosis of depression .......................................25
2.7 Summary ...............................................................................................................31

CHAPTER THREE: ..........................................................................................................32
3.1 Preview ..................................................................................................................32
3.2 The ORYGEN Y-H Database Collection Procedure ........................................33
   3.2.1 Stage T1 of Data Collection ....................................................................33
   3.2.2 Stage T2 of Data Collection ....................................................................35
3.3 ORYGEN Y-H Database Recording and Annotation .........................................36
   3.3.1 Audio-visual recordings of family interactions .........................................36
   3.3.2 Observational annotation of family interactions .......................................39
3.4 How the ORYGEN Y-H Data Facilitated Prediction of Depression? ...............40
3.5 Summary ..............................................................................................................41

CHAPTER FOUR: .............................................................................................................43
4.1 Preview ..................................................................................................................43
4.2 Image Corpus .......................................................................................................44
4.3 Image Based Depression Prediction Algorithm ..................................................46
4.3.1 Algorithm overview.................................................................46
4.3.2 Image pre-processing...............................................................47
  4.3.2.1 Image normalization..........................................................47
  4.3.2.2 Frontal face detection using local Successive Mean Quantization
            Transform (SMQT) features and Split up Sparse Network of
            Winnows (SNOW) classifier......................................................48
4.3.3 Facial image feature extraction ..............................................49
  4.3.3.1 Eigenfaces (PCA)..............................................................51
  4.3.3.2 Fisherfaces (PCA+LDA).......................................................53
4.3.4 Image modeling and classification.........................................54

4.4 Experiments for Image Based Depression Prediction..................56
  4.4.1 Experimental setup..............................................................56
  4.4.2 Person independent classification results...............................58
  4.4.3 Person dependent classification results.................................60
    4.4.3.1 Selection of an optimal decision rule for the person dependent
            classification.................................................................60
    4.4.3.2 Results of the person dependent classification using the optimal
            classification decision rule..................................................62

4.5 Summary and Conclusions.......................................................64

CHAPTER FIVE:.................................................................................67
  5.1 Preview......................................................................................67
  5.2 Speech Corpus............................................................................70
  5.3 Speech Pre-processing..............................................................73
  5.4 Calculation of Acoustic Features.............................................75
    5.4.1 Glottal features (G):............................................................77
      5.4.1.1 Time domain glottal parameters (G-T)..............................79
      5.4.1.2 Frequency domain glottal parameters (G-F)......................83
    5.4.2 Prosodic features derived from the speech wave (PS)............85
      5.4.2.1 Fundamental frequency (PS-F0)......................................86
      5.4.2.2 Logarithmic Energy (PS-LogE)........................................86
      5.4.2.3 Jitter (PS-J).................................................................86
      5.4.2.4 Shimmer (PS-S)............................................................87
    5.4.3 Prosodic features derived from the glottal waveform (PG)......87
    5.4.4 Teager Energy Operator features (TEO)...............................88
      5.4.4.1 Teager Energy Operator based features derived from speech
            waveform (TEOS)..............................................................89
      5.4.4.2 Teager Energy Operator based features derived from glottal
            waveform (TEOG)..............................................................91
    5.4.5 Spectral features derived from speech waveform (SS)............94
      5.4.5.1 Spectral flux.................................................................95
      5.4.5.2 Spectral controid..........................................................95
      5.4.5.3 Spectral entropy............................................................96
      5.4.5.4 Spectral roll-off............................................................96
      5.4.5.5 Power spectral density (PSD)..........................................96
      5.4.5.6 Formants.................................................................96
  5.5 Gaussian Mixture Model (GMM) and Bayesian Classification ....97
# LIST OF TABLES

Table 4.1: IMAGE CORPUS OF “AT RISK” (AR) PARTICIPANTS ........................................ 45

Table 4.2: PERSON INDEPENDENT CLASSIFICATION PERFORMANCE RESULT ............... 59

Table 4.3: CLASSIFICATION DECISION RULES ............................................................................. 62

Table 4.4: PERSON DEPENDENT CLASSIFICATION PERFORMANCE RESULT ..................... 64

Table 5.1: SPEECH CORPUS OF “AT RISK” (AR) PARTICIPANTS ........................................ 73

Table 5.2: SUMMARY OF ACOUSTIC FEATURE CATEGORIES .............................................. 76

Table 5.3: CRITICAL BANDS USED TO CALCULATE THE TEOS FEATURES ......................... 91

Table 5.4: CRITICAL BANDS’ (CB’) GROUPS, CB’ CENTER FREQUENCIES AND
     BANDWIDTHS USED TO DETERMINE THE TEOG FEATURES ....................................... 94

Table 5.5: DEPRESSION PREDICTION RESULTS FOR INDIVIDUAL FEATURES FROM
     SPEECH WAVEFORM USING THE UTTERANCE BASED (UB) APPROACH ...................... 109

Table 5.6: DEPRESSION PREDICTION RESULTS FOR INDIVIDUAL FEATURES FROM
     SPEECH WAVEFORM USING THE PERSON BASED (PB) APPROACH .......................... 109

Table 5.7: DEPRESSION PREDICTION RESULTS FOR INDIVIDUAL TEOG SUB-BAND
     FEATURES USING THE PERSON BASED (PB) APPROACH ............................................. 112

Table 5.8: DEPRESSION PREDICTION RESULTS FOR THE MCSC SYSTEM WITH FOUR
     CHANNELS (G, PS, TEOS & SS) USING THE PERSON BASED (PB) APPROACH ....... 118

Table 5.9: DEPRESSION PREDICTION RESULTS FOR THE 1W AND 2W MCWSC METHOD USING THE UTTERANCE BASED (UB) APPROACH ........................................... 127

Table 5.10: DEPRESSION PREDICTION RESULTS FOR THE 1W AND 2W MCWSC METHOD USING THE PERSON BASED (PB) APPROACH ............................................. 128

Table 5.11: OMCWSC DECISION MATRIX FOR A SINGLE SET OF WEIGHTS (1W) .............. 132

Table 5.12: OMCWSC DECISION MATRIX FOR A TWO SET OF WEIGHTS (2W) ............... 132
Table 5.17: AN EXAMPLE OF PERMUTATION MATRIX FOR 3 CHANNELS (M=3) .......... 136

Table 5.18: DEPRESSION PREDICTION RESULTS FOR THE OMCWSC SYSTEM WITH TEOG-SB1-SB5 (100Hz – 5500Hz) FEATURES ................................................. 139

Table 5.19: DEPRESSION PREDICTION RESULTS FOR THE OMCWSC SYSTEM WITH TEOG-SB2-SB5 (700Hz - 5500Hz) FEATURES ........................................ 139

Table 5.20: DEPRESSION PREDICTION RESULTS FOR THE OMCWSC SYSTEM WITH TEOG-SB3-SB5 (1300Hz – 5500Hz) FEATURES ...................................... 139

Table 5.21: DEPRESSION PREDICTION RESULTS FOR THE OMCWSC WITH SELECTED FEATURES (G, PS AND TEOG (SB3-SB5)). .................................................. 142

Table A1: Representation of the codes used in the LIFE coding system ............ 150
LIST OF FIGURES AND ILLUSTRATIONS

Figure 3.1: Longitudinal data collection procedure ......................................................... 33
Figure 3.2: Participant selection procedure in stage T1 ................................................... 38
Figure 4.1: Block diagram of the proposed image based prediction methodology .......... 46
Figure 4.2: Examples of images from the publically available FERET database
    generated using the eigenface (PCA) method. (a) Frontal face images (b)
    Calculated mean face of a dataset. (c) Examples of eigenfaces for the dataset in
    (a). ............................................................................................................................. 52
Figure 4.3: Overall modelling and classification for the training and testing phase ....... 56
Figure 4.4 ROC space analysis ......................................................................................... 63
Figure 5.1: A general framework for the speech based prediction of depression......... 70
Figure 5.2: Examples of a speech frame (a), glottal wave (b), first derivative of the
glottal wave (c). The parameters marked on these pictures were used to derive
the 9 G-T features. .................................................................................................... 83
Figure 5.3: An example of the glottal wave spectrum. H1 and H2 denote the
magnitude of the first and second harmonics respectively ........................................ 84
Figure 5.4: An example of the TEO-CB-Auto-Env feature extraction within the 6th
critical band for a single speech frame: (a) TEO energy profile (b)
Autocorrelation of the energy profile ........................................................................ 90
Figure 5.5: Example of the TEOG feature extraction within the 9th critical band for a
single glottal frame: (a) TEOG energy profile (b) Autocorrelation of the energy
profile ........................................................................................................................ 92
Figure 5.6: TEOG feature extraction implementation ...................................................... 93
Figure 5.7: An overview of the training stage of single-channel system generating AR
and NAR models for a given type of features (G, P, TEO or S) .............................. 107
Figure 5.8: An overview of a single-channel classification system using models from
the training stage ......................................................................................................... 107
Figure 5.9: Overview of the two-stage GMM multi-channel classification system ...... 117
Figure 5.10: Determining the weight values for the MCWSC classification decision... 122
Figure 5.11: The multi-channel classification system using MCWSC system weights . 124
Figure 5.12: Optimized multi-channel weighted speech classification (OMCWSC) system testing stage. 134

Figure 5.13: Training process for finding the optimal weight vector in the OMCWSC approach. 135
Chapter One:

INTRODUCTION

1.1 Preview
This chapter explains the problems related to the large prevalence of depression cases in adolescent populations. The increasing need for development of efficient methods for an early detection of depression symptoms is discussed leading to the problem statements for this thesis. This is followed by sections explaining the thesis aims, thesis scope and limitations, thesis contributions and the overall structure of the thesis chapters.

1.2 Background and Problem Statement
Over the recent years, the practice of delivering mental health services has changed dramatically. Medical and psychological assessment is being automated step-by-step and consultation is increasingly conducted via the Internet.

Given the prevalence of mental health disorders and the burden on economies worldwide, mental health informatics become a dedicated branch of engineering and applied computer science. Automated, easy to use mental health diagnostic and prediction techniques are sought after for online and mobile phone applications. Mobile phones can record and transmit speech, text and movement data. They can store and
communicate personal information, including medical diagnostic information. For the first time in history, objective data about the behavior and mental health of hundreds of millions of individuals can be recorded, stored and analyzed. With the help of dedicated computer software, it also provides a new unique opportunity to predict and prevent mental health problems within large populations.

Speech and facial image are particularly interesting from the viewpoint of psychological assessment. For instance, depression may change the characteristics of voice and details of facial expressions in individuals and these changes can be detected by a special form of speech and/or image analysis. Computational screening methods that utilize speech and image can detect subtle changes and alert clinicians as well as individuals and caregivers [70].

Depression is one of the most common medical illnesses which affect people of all ages and gender. Sufferers from this psychological disorder are often linked to experiencing various negative effects such as depressed mood, decreased motivation and energy, loss of appetite and low self-esteem which impairs a person’s ability to carry out daily activities normally. If left undetected or not treated, it can seriously jeopardize an individual’s social, emotional, educational and vocational capabilities. At its worst, depression can also lead to suicide where an increase in the number of suicide attempts and deaths among adolescents has been a major concern [65], [136].

At the beginning of the 21st century, the World Health Organization estimated that 121 million people in the world suffer from clinical depression [21]. Moreover, depression is estimated to become the world’s second-greatest burden of disease by the year 2020.
In many cases, the life-long reoccurrences of depression during adulthood are consequences of untreated or undetected depression symptoms appearing for the first time during childhood or adolescence [90]. The onset of depression is most likely between 13 and 17 years of age, with Major Depression (MD) being the most common form. An early intervention, preventing the onset of clinical depression would provide a very important method for reducing the burden of the disease. Efficient depression prediction techniques are therefore needed to determine the risk for an onset in young individuals.

The diverse nature of depression leads to numerous research challenges, including the understanding of its causes and effects, as well as development of efficient treatments and prevention methods.

Medical examinations and psychologist analysis, based on factors such as behavioral, genetic, psychological and physiological changes are good indicators of detecting depression at its existence. However, in order to detect the possibility of an occurrence of a depression episode in the near future at the point where no symptoms of depression were detectable through conventional medical and psychological diagnosis is a major challenge. Probability tests for conventional methods and combination methods with other area of study were used in order to identify the risk of the disorder. For example, common psychological assessments uses methods such as self-reports and depression scale ratings for the prediction of depression in adolescents [1]. Some combine studies include medical imagery information analysis by investigating the correlation of brain structure with depression symptoms [150].
Although depressive disorders are more often a problem in adults, report has shown that at least 14% of adolescents aged 4-12 years were diagnosed with depression or showed some signs of mental disorder. This figure rose to 19% for the 13-17 years age-group [118]. Depressions in young people are often not recognized, with a general perception that it is uncommon for suffers at such a young age.

The risk of allowing such psychological problem to develop in adolescents are not only limited to issues relating to reduced social and vocational opportunities but can also lead to suicidal cases. Thus, it is important that early predictive signs of depression are detected so that preventative measures can be applied before a full-blown disorder becomes evident.

This work focuses on developing an effective prediction system that is able to determine in objective terms, if an adolescent is at possible risk of clinical depression in the near future. The objective character of the proposed systems comes from the fact that the prediction is based on the analysis of images and acoustic speech parameters. Just like a blood analysis can determine if a patient suffers from a particular disease, the analysis of images or speech sounds can provide vital diagnostic information facilitation assessment of the patient’s mental health.

Aside from containing important diagnostic information, speech and image signals are very easy to capture in a discrete and non-invasive way. The speech or image based diagnosis is a low cost procedure and can be conducted in a fully automatic way by a person that does not necessarily have an extensive experience in mental health care. These attributes make the proposed system particularly suitable as a mass screening device for detecting “at risk” populations of children.
The proposed methodologies can be also extended to testing of war veterans, elderly people, young mothers and other groups particularly vulnerable for depression.

1.3 Thesis Aims

The study was motivated by an existing need for an easy to use on a mass scale, objective methods for an early prediction of risk for depression.

The thesis aimed to provide answers to the following major research questions:

1) Is it possible to predict the risk for depression from facial image features 2.5 years before the full blown symptoms occur?

2) Is it possible to predict the risk for depression from acoustic speech parameters 2.5 years before the full blown symptoms occur?

3) What kind of facial features and acoustic speech parameters provide the best discrimination between the “At Risk” (AR) and “Not AT RISK” (NAR) adolescents?

4) What is the most efficient image or speech modeling and classification procedure for an early prediction of depression?

1.4 Thesis Scope and Limitations

This thesis has a unique aspect, providing forerunning investigation into the prediction of clinical depression in adolescents (9-12 years of age) based on facial images and acoustic speech characteristics.

The prediction framework investigated in this thesis is based on automatic image and speech classification into two classes (binary classification) : “at risk” and “not at
risk” of developing depression in the near future, which means within 2.5 years from the
time the analysed speech recordings were made.

This study was facilitated by the specific character of the available data base
containing audiovisual recordings of participants who were “at risk” and “not at risk” of
developing depression. Since the data collection procedure was not originally designed
for this particular study but for other types of psychological studies of risk for depression
in adolescents, the nature of the available data imposed certain limitations to the scope of
this thesis.

These limitations included:

1) The tested data contained 15 (6 male and 9 female) “at risk” individuals and 15
(6 male and 9 female) “not at risk” individuals. This may appear to be a
relatively small sample in terms of individuals (but not in terms of length of
recordings; 20 minutes/person for each session) however, as explained in
Chapter 3, these numbers resulted from the fact that only 15 individuals out of
191 initially healthy participants developed Major Depression during the course
of data collection.

2) The data base contained only speech of adolescents (9-12 years old); therefore
no other age groups could be tested. However, the age represented by the data
base is a typical age when the onset of depression appears. It is therefore very
important to study this age group for the purpose of mass screening and early
prevention of depression.

3) Although, the depression disorder covers a wide range of different types of
depression, the prediction methodology presented in this thesis was tested only
in the case of Major Depression (MD). The MD is the most frequent type of depressive disorder found in adolescents and it happened that this particular type of depression was detected in the 15 individuals participating in data collection. As explained in Section 3, other types of depression were also detected in some participants but the numbers were too small to conduct useful tests.

4) The tests were limited to English speaking participants selected from a population of Melbourne school students.

1.5 Thesis Contributions
This thesis provided important contributions to the research of methods for an early prediction of depression in adolescents.

The study had a very unique character and pioneering character. To our knowledge it was the first study of its kind.

The prediction methodologies investigated in this thesis was based on an automatic image analysis and classification of facial images and speech acoustic. The proposed methodology had an objective and quantitative character which made it particularly suitable for applications with young children who may not always be mature enough to participate in standard depression screening methods based on questionnaires.

Most importantly the thesis provided answers to the major research questions and developed new, efficient methods using acoustic speech parameters and optimised multi-channel classification system that is capable of predicting the risk of depression 2.5 years before. This can be summarised as follows:
1) It was shown that classification of facial images is generally less efficient in prediction of depression than the classification of acoustic speech parameters;

2) It was demonstrated, that it is possible to efficiently predict the risk for depression from acoustic speech parameters 2.5 years before the full blown symptoms occur;

3) Statistical and classification tests conducted on a wide range of speech parameters, revealed which parameters provide the most efficient discrimination between adolescent who are “at risk” and “not at risk” for depression;

4) The study proposed a new multi-channel weighted speech classification (MCWSC) approach. It was found that the optimised version of this method, called the optimised multi-channel weighted speech classification (OMCWSC) system is the most efficient speech based method for prediction of depression in adolescents.

1.6 Thesis Structure

The thesis chapters are organized as follows:

   **Chapter Two** explains the problem of clinical depression in adolescents and its impact on society. This is followed by review of existing studies related to depression recognition from speech and facial images. Different types of features, modelling techniques and general computational frameworks are discussed and compared.

   **Chapter Three** describes the database of audiovisual recordings used in experiments. The process of data collection and its importance to the prediction setup used in this study is explained.
Chapter Four describes methodology and experiments used to test the image based prediction of depression. Processing stages including facial image corpus, pre-processing, feature extraction methods and classification tasks are described. Experimental results are presented and discussed.

Chapter Five describes methodologies and experiments used to test the speech based prediction of depression. First, the formulation of the speech corpus is explained. Next, the classical single-channel classification methods and experimental results are presented. This is followed by development and testing of new multi-channel classification approaches.

Chapter Six provides research summary and suggestions for future work.
Chapter Two:

PREVIOUS STUDIES OF DEPRESSION RECOGNITION FROM

SPEECH AND FACIAL IMAGE

2.1 Preview

This chapter provides an overview of previous studies aiming to understand the relation between clinical depression and vocal indicators or facial expressions. Emotions can be revealed in many different ways, where speech and facial expressions are one of the most common sources of approach in expressing emotions. This strong correlation between speech or facial expressions in emotions and depression are significant as it paves a link between each element. The concept of automatic depression recognition through acoustic and facial analysis channels has been widely researched due to the need and complexity of these studies. The following sections in this chapter provide a general understanding of the concepts given by past researchers that bare a close relation to the subject of this study. This literature review provides background knowledge of existing methodologies in speech and facial image signal processing correlates of clinical depression. The difference between detection of depression and the prediction of an onset of depression is also discussed here, as most studies using vocal speech and facial image sources focus on the detection tasks when symptoms of depression are already present. Investigations into
depression tendencies (before the full blown symptoms are present) are still limited to other modes of psychological assessments.

2.2 Clinical Depression
Clinical depression belongs to the group of affective (mood) disorders in which emotional disturbances consist of prolonged periods of excessive sadness. Depressed individuals suffer from varying degrees of psychomotor retardation or agitation. In terms of emotional state, depression sufferers often experience lasting feelings of hopelessness, anger, guilt, desperation and loneliness which often leading to suicidal thoughts. The specific emotional state of a person suffering from clinical depression affects facial expression as well as the acoustic qualities of the voiced speech. Symptoms of depression can be therefore detected through an analysis of facial images and speech characteristics.

The emotional health of individuals has a significant impact on national health and economic outcomes [92]. Emotional disorders, such as anxiety and depression, are amongst the most common and disabling illnesses of any type observed in the worldwide community, and have been recognised as some of the major public health issues.

Existing diagnostic and treatment resources are very limited and likely to remain this way in the foreseeable future. Therefore, it is particularly important to look for new preventative measures that reduce the impact of depression on humanity.

Depressive disorders seriously affect social, emotional, educational and vocational outcomes. It is also the most common precursor of suicide. It is estimated that up to one in eight individuals will require treatment for depressive illness in their lifetime.
The prevalence of depression, the world's fourth most serious health threat in 1990, is expected to rise steadily.

Depression and anxiety account for most of the economic, social and personal costs of mental disorders. Depression greatly impairs a person's ability to function physically, socially, and at work. In adolescents depression may be associated with loss of energy and social withdrawal but may also result in disruptive behaviours or substance use (drugs and alcohol). In the longer term, depression can reduce social and vocational opportunities for young people as a result of early school dropout and sporadic employment opportunities.

2.3 Depression in Adolescents

The onset of depression is most likely at age 18, with Major Depression (MD) being the most common form. An early intervention preventing the onset of clinical depression would provide a very important method for reducing the burden of the disease. Efficient depression prediction methods are therefore needed to determine the risk for an onset in young individuals. A report from the Institute of Medicine of the United States in 1994 [91], strongly indicates the need to increase research on preventive methods in order to reduce the incidence (that is, the number of new cases) of mental disorders in the population. One of the goals for a prevention research is the clear identification of groups at high imminent risk for the disorder.

Depression appears at different stages of the life cycle and it is one of the most common mental health problems in adolescents. Adolescents are usually defined as aged 9-20 years. It is estimated that a depressive episode affects between 14-30% of young
females and 13-17% of males [27], [122]. Depression also occurs in children younger than 13 years of age, but becomes progressively more prevalent after puberty and reaches adult levels in the late teens. During the early stages depression is often unrecognised and under-treated [109], leading to many undesirable social, health, education and employment issues.

Early diagnosis of depression is extremely useful and can mean a minimal disturbance of typical functioning and development of social and academic skills. Unfortunately children often are not be able or not willing to express verbally their feelings and thus help with an early diagnosis. Also, adolescents are more unlikely to understand or determine if they are in a depression state compared to adults where professional help are often not looked for. The conventional diagnosis of depression in adolescents is often based on observations of behavioural patterns, and interviews with parents and teachers. This process is time consuming and the illness is usually recognised when in advanced stages.

Current diagnostic methods are qualitative and largely based on the personal skills, experience and intuitive judgment of a mental health practitioner. The number of highly skilled professionals is limited, and their availability is restricted to major towns and health centres. As a result, each year, thousands of cases of depression are not being diagnosed and left untreated, leading to potential suicides. These problems are of particular concern in the rural areas of many countries where the number of suicides amongst young people is alarmingly large.

An automatic speech analysis in diagnosis of depression is currently an important subject of ongoing research. It can provide an efficient and reliable mass
screening methodology for detection of depression in adolescents and such addressing the huge prevalence of the depression problem in adolescents.

2.4 Detection Versus Prediction

Depression is known as one of the most common psychological condition that affects people of all ages. Kessler (2003) reported that the age group of 18-29 year olds are 70% more likely to have already experienced an episode of depression over their lifetime and the figure increased to 120% more likely when compare to adults above the age of 60 years [63]. The need for early prevention and interventions has been even more crucial with the number of reported cases rapidly increasing and it is most likely that people who experience a first onset of depression episode will encounter recurrences. An onset of depression is most likely to occur in early adolescence stage to late adolescence and most adults who experience depressive recurrences would have experience their first symptoms during this adolescence stage [88], [107]. A strong focus on adolescents’ mental health issues is needed with the diverse family, school and work related challenges faced by affected young people and the negative effects being often prolonged towards adulthood [105].

The common distinction between detection task and prediction task or between treatment and prevention can be generalized as follows; if at the point of diagnosis a depression episode was currently or previously present, any related approach would be considered a detection or treatment task. Whereas, if at the point of diagnosis, there was no current detectable depression episode and there was no evidence of previous episodes, then any related approach would be considered a prediction or prevention task. This means that during a prediction or prevention study, participants were all clinically
diagnosed as having no symptoms of depression at the time where the study was conducted.

Majority of previous studies were focused on the detection and treatment of already existing depression. Prevention studies searching for possibilities of preventive measures in order to avoid the onset of a depression are still in the investigating stages. Similarly, the clinical practice is also limited to detecting or recognizing existing episodes of depression where sufferers are already bounded by the psychological disorder and clinical treatments are already needed. The use of depression rating scales is one of the most common measures and guidelines in clinical practices to make a conclusive diagnosis of clinical depression based on different categorical assessments, symptoms and factors analyzed. Some of the common rating scales used include the Hamilton Depression Rating Scale [46], Center for Epidemiologic Studies Depression Scales [30], [111], [114], and Beck Depression Inventory [13], [114].

Given the increasing need for efficient depression prediction methodologies for adolescents, recent studies’ using psychological assessments has investigated the possibility of predicting future depression episodes in adolescents at the point where no depression symptoms are detectable [68], [80], [140], [141]. Preventative research has been focused on risk factors associated with depression and intervention tasks [87], [88], [89], [90], [91]. Some of the areas of research relating to assessing risk factors associated with depression tendencies include the studies of genetic epidemiology of depression [29], [130], parenting influence on adolescent depression [31], [52], [150], and cognitive styles in the likelihood of depression development [137]. A combination of several putative risk factors was also considered in predicting major depression in adolescents
Examples of most often used prediction approaches include studies of behavioral and psychological states of individuals \cite{90, 114, 149} and identification of characteristics such as family history, early adversity, gender, age or socioeconomic status \cite{90}. Numerous studies have undertaken the task of recognizing various early behavioral and psychological signs and symptoms of the disorder that are not classified as full-blown symptoms \cite{87, 89, 91}. For example, children of parents being treated for depression who score high in depression symptoms scales are at higher risk of developing clinical depression during adolescence \cite{42}. Similarly, women who are pregnant and have high depression symptoms levels are at high risk of developing pre- and post-partum depression \cite{52}. Individuals with one or two short alleles of the serotonin transporter gene are more likely to develop a clinical depression when they are faced with stressful life events \cite{29}. In \cite{67, 68}, the concept of emotional inertia characterizing a specific type of psychological maladjustment showing high resistance to emotional changes was investigated. People with high emotional inertia are resistant to external influences and preserve their current emotional state for a long time. It was found that depressed individuals in particular tend to have higher levels of emotional inertia. In \cite{68} emotional inertia was shown to predict the emergence of clinical depression in adolescents 2.5 years later.

2.5 Depression Recognition from Facial Images

Facial expressions correlate with the self-reported experience of emotion and are the most important non-verbal (observable) indicators of emotions. From the psychological perspective, these observable emotional expressions are an influential guide in
understanding behavioural changes and mental states. The disturbance in emotion expression is seen as a central issue in the aetiology of depression, therefore measurement of facial expressions could be useful in clinical investigations and in the treatment of depression. Patients with depression [33] show a lot more sadness and anger than people who are not depressed. Facial behaviour during depression shows lack of emotional verability [113]. In [43], it was observed that facial muscle activity over the brow and cheek region was reduced in depressed compared with non-depressed patients indicating certain psychomotor retardation associated with depression.

One of the disadvantages of using facial images for clinical detection and prediction of depression is that, image analysis methods present extremely extensive computational task. In most cases, complex analytical procedures need to be employed to provide an adequate representation of various facial regions. This task in combination with the large numbers of images required for the purpose of system training makes the whole process quite cumbersome.

2.5.1 Representation of Facial Expressions

As facial expressions can represent different emotions, moods and feelings, facial image analysis methods are suitable for applications such as: person identification for security systems, detection of pain in patients, assessment of mental states and psychopathology studies. For these, as well as many other applications, it is important to understand the relation between facial expressions and the corresponding facial movements or other changes in facial appearance. An analysis of these changes represented in different ways depending on the undertaken approach, provides information necessarily to solve a given
image analysis task. Ekman and Friesen in the 1970s introduced a coding system representing the facial behaviours known as the Facial Action Coding System (FACS) [32]. An observed expression was coded based on measurement units known as Action Units (AUs) representing different parts and movements of the facial region [34]. The changes in facial expressions in relation to the corresponding emotions and mental states can be represented in many different ways. For example, the appearance based facial representation is a popular approach that generates texture features which best represents a facial image. Some of the popular approaches used in the facial image analysis are the Principal Component Analysis (PCA) [138], Linear Discriminant Analysis (LDA) [14] and Independent Component Analysis (ICA) [12]. Template or feature based representation of facial image is an approach which uses geometric facial regions or template modeling of the facial region to represent a facial characteristic or appearance. Examples of methods using this approach include the Active Appearance Model (AAM) [22] and Gabor features [129].

2.5.2 Facial Image Analysis and Classification

The study of facial images relating to depression varies in terms of its implementations, techniques and applications. Expressed emotions are closely related to behavioral changes with some emotions or moods are more dominant when a person is suffering from depression compared to mentally control person [57], [115].

Psychological assessment which tests the emotional reactivity of sufferers from clinical depression uses facial images representing moods and emotions as stimuli for emotion recognition tasks. This kind of study assesses a person’s ability to identify
different types of emotional expressions form facial images presented and to determine the deficits in facial emotion recognition of the participants [81], [132], [133]. Surguladze et al, identified some inconsistency of studies relating to the impaired identification of facial emotion in depressed patients where some studies are for and against the capability of using visual perceptual deficit for this purpose [132].

Some of the early work analyzes facial expressions of patients with depression and comparing with facial expressions of control patients. The difference and changes in patterns of facial expressions exhibit by the two groups provided the discriminative basis between individuals who are experiencing a depression episode and individuals of stable state [15], [33], [34], [41], [43].

The use of facial image analysis in pattern recognition task for depression detection in individuals are still limited in contrast to the number of automatic detection of depression studies conducted on speech, partially attributed to the fact that image analysis is generally computationally more demanding. However the use of facial images for automatic depression recognition has attracted researches in recent times to investigate if facial parameters and expressions carry important distinctive information for depression diagnosis.

The investigation of using facial image based recognition of depressed and non-depressed adolescents was performed by Maddage et al. [77]. A small dataset containing 4 subjects clinically diagnosed as depressed and 4 control subjects were selected for the classification task. The experiments used Gabor wavelet features extracted at facial landmarks together with Gaussian Mixture Model (GMM) classifier, reported up to
85.5% of correct classification for recognising depressed female subjects and 87.7% of correct classification for recognising non-depressed female subjects.

Cohn et al. [20] investigated on using manual facial action coding system (FACS) annotation and also the active appearance modeling (AAM) method as facial image features together with the Support Vector Machine (SVM) for a classification task of detecting depressed and control participants. An overall classification accuracy of 88% was achieved using the manual FACS annotation and 79% accuracy using the automatic AAM facial feature classification. A similar approach was also conducted in the studies of McIntyre [79], where Region Units (RU) and active appearance modeling (AAM) were used to measure facial activity in depressed subjects.

The work on using FACS coding was extended by Girard et al. [44], where manual and automatic FACS annotations were compared to analyze facial expressions of depressed patients and investigate the relationship between the change in severity of depression and facial expression. It was found that patients in highly depressed state exhibit different patterns of facial expression activity where more facial expressions associated with contempt was shown, compared to low symptom depressed state patients. It was also justified that the automatic FACS was consistent with the manual FACS annotation producing the same pattern of depression effects. Localized Gabor features was used together with support vector machine (SVM) classifier for the automatic FACS implementation.

Valstar et al. [139], performed two recognition task using video and audio features relating to a depression study. First, continuous emotion recognition based on the affective dimensions valance and arousal was investigated and, second, to predict a single
depression indicator for the depression recognition task. Local Phase Quantisation (LPQ) histograms were used to represent facial image features to perform the two recognition tasks. A mean absolute error (MAE) of 10.88 and root mean square error (RMSE) of 13.61 was reported for the depression recognition from facial images.

Some significant study on the use of facial images for the detection of depression was seen in [58], [60], where the Space-Time Interest Points (STIP) and Local Binary Patterns Three Orthogonal Planes (LBP-TOP) features were investigated. In a more recent work, Joshi et al. [59], compared a unimodal and multimodal system for depression detection using these video features together with audio features. The two types of video features, (STIP) and (LBP-TOP) using a non-linear SVM classifier were compared. Features extracted using the STIP concept is computed over unaligned raw video frames where key interest points are detected by video level clustering. For the LBP-TOP features, facial images were first detected and aligned, and then the LBP-TOP descriptors were assigned to capture the intra-face movements. The experimental results on the video features achieved classification accuracy of up to 76.7% with the STIP features alone, and up to 81.7% classification accuracy with a STIP + LBP-TOP feature combination.

The study of using facial image analysis for depression detection has shown promising results and is feasible for an automatic depression analysis. However, all these research were conducted on facial images of subjects which were currently or previously medically diagnosed with depression. It is expected that there are more distinguishable differences in facial features between subjects that has depression experience than those without. Preliminary test on facial images of subjects that have not yet been diagnosed
with depression using appearance based features (PCA and LDA) were experimented in this thesis for the purpose of comparing the feasibility of using facial image features to acoustic speech features. As there are subtle differences to compare between subjects of the same class (only between subjects not yet been diagnosed with depression) in a prediction task compared to a detection task (between depressed and non depressed subjects), the appearance based features which are based on the overall facial section rather than sections or landmarks of the face should provide a more holistic analysis for a prefatory measure.

2.6 Depression Recognition from Speech

This section focuses on the verbal evaluation in speech communication, emotional states and the link with clinical depression. Acoustical correlates of depression have been a research area of interest as the vocal auditory production is regarded as one of the most important connection with the physiology of emotions. Careful analysis of the acoustic characteristics would provide useful information of certain pattern or traits associated with depression.

2.6.1 Physiological basis for speech based diagnosis

Studies conducted since the 1980’s (Scherer, 1979, France, 1997, Ozdas, 2001), gathered considerable evidence that emotional arousal produces changes in speech production by affecting respiratory, phonatory, and articulatory processes. This is largely due to the fact that emotional arousal activates the somatic nervous system and the sympathetic and parasympathetic autonomic nervous system (ANS). For example, the temporal rate of
articulation and the frequency range of vowels can be modified by changes in the timing of the muscle movements controlling the motion amplitudes of the articulatory structures.

These muscles are driven by the limbic control system. The sympathetic and parasympathetic effects controlled by the limbic system can influence speech production by introducing changes in respiration and phonation. Resulting changes in subglottal air pressure and vocal cord tension could have an effect on the fundamental frequency of vocal cord vibration (F0). Other ANS-dependent effects such as mucus secretion may alter speech characteristics during intense emotional arousal. There is evidence that psychomotor disturbances are particularly distinctive symptoms of mood disorders. Symptoms of psychomotor agitation or retardation may actually precede the full onset of a depressive episode and thus provide early indication of high risk for mental illness.

2.6.2 Speech acoustics as correlates of Depression

Emotional excitation through speech was investigated by Scherer in 1979 [119], where the changes in speech production mechanism is related to the vocal expressions affected by the emotional stimulation. It was also found that expressive emotions in speech are strong correlates of depression [78]. The analysis of acoustic characteristics relating to emotional arousal that influences the speech production started with the investigation of changes in the fundamental frequency by Scherer [120], [121]. Emotion recognition in speech has been extensively researched over the past years with the advancement of different methodologies in characterizing acoustic parameters such as pitch, Teager energy operator and vocal tract features [95], [96], [97], [126], [127], [142], [143].
The use of acoustic parameters as indicators of psychological stress were also effective, as stress related changes affects the speech production and thus the vocal outcome. Studies have shown the effectiveness of using acoustical features in recognizing speech under stress [47], [97], [148], where more recent studies reporting classification accuracy of more than 90% using the non linear Teager Energy Operator features [48], [51], [98]. With the strong relation between emotional stress and clinical depression, where stress is one of the factors leading to onset depression [83] and the proven capability of acoustic features in recognizing stress, there has been a strong link between acoustical parameters and clinical depression.

This observation led to numerous studies searching for acoustic correlates of depression (France et al., 1997, Ozdas, et al., 2004, Moore et al., 2008, Low et al., 2011). The most often investigated parameters included prosodic parameters such as fundamental frequency (F0), formants, jitter, shimmer, intensity of the speech signal, and speech rate. Other commonly used speech parameters include cepstral features (i.e. mel frequency cepstral coefficients), spectral features (i.e. power spectral density) and glottal features. The Teager energy operator (TEO), which measures the number of additional harmonics due to the nonlinear airflow in the vocal tract that produces the speech signal also, attracted attention of researchers. There is strong evidence that these parameters could be used to discriminate between non-depressed and depressed adult patients, and possibly between different stages of depressive disorder [20], [39], [84], [85], [104].

Acoustic speech parameters can be applied in the process of depression diagnosis and also as indicators of effectiveness of various treatments for depression. Due to significant differences between adult and adolescent speech [109], studies specifically
designed to investigate the effects of depression on speech in adolescents are needed. This should provide vital information for the design of diagnostic algorithms aiming to detect onset of depression and assess the risk of depression in adolescents.

2.6.3 Studies of speech based diagnosis of depression

The link between emotional excitation, changes in speech production and vocal expressions have long been investigated by psychiatrists and clinical psychologists. A subjective assessment that differences in the characteristics of patients’ voices can be used to perceive as indicators of their mental states.

The level of correlation built to recognize complex relationships between speech parameters and depression has customarily been assessed using multivariate analyses. Kuny et al. [66], investigated the recovery progress of depressive patients by assessing the changes in speaking behaviour and voice characteristics. These voice changes provided an observational distinction for which characteristics were more dominant or diminishing during the course of recovery from a depressive state. Ellgring and Scherer [35], compared several voice and speech parameters of patients during depressed and recovered mood states in a longitudinal study. The results showed that an increase in speech rate and a decrease in pause duration are powerful indicators of mood improvement. It was also observed that only in female patients, there was a decrease in minimum fundamental frequency of the voice predicted mood improvement. In Alpert et al. [9], the study focuses on measures of fluency such as speech productivity and pausing, and prosody which includes emphasis and inflection. Observations shown less prosody were portrayed in depressed patients. Fluency measures were also reported providing
different trends in the depressed patients. It was concluded that acoustics measures may provide an objective evaluation of depression.

The research of acoustical characteristics associated with depression and classification task of sufferers from major depression and control groups was on going with studies relating to acoustic correlates of emotions and stress. Scherer investigated the non linguistic category of speech such as prosody to understand the emotional and psychological state of a person [119], [120], [121]. Some of the commonly used acoustic parameters in indentifying the depressive related cues or symptoms in human speech include spectral category features, mel frequency cepstral coefficients, glottal features and Teager energy operator features.

Application of such acoustical features for the purpose of detection of depression in individuals was significant. Studies have demonstrated that acoustic speech analysis can be efficiently used to recognize symptoms of depression in adults as well as adolescent. Investigation by France et al. [39], analyzed the fundamental frequency, amplitude modulation, formants and power spectral density (PSD) of speech samples from a total of 34 control subjects and 42 major depressed patients. The formants and PDS features provided the best classification accuracy of up to 94%. Ozdas et al. [104] experimented on the vocal jitter and glottal spectrum features from speech samples of 30 subjects provided an overall classification accuracy of 90% for the depressed and control classification task for the combined jitter and glottal spectrum feature.

Some of the more current research to date includes studies from Moore et al. [84], [85], Cohn et al. [20], Low et al. [73], [74], [75], Alghowinem et al. [3], [4], and Scherer et al. [123]. Moore et al., in his investigation compared the deviation of prosodic
and glottal feature statistics of speech samples from depressed patients. With a data
csample size of 33 subjects, a binary classification into depressed and control categories
were examined using a feature selection strategy of quadratic discriminant analysis. It
was observed that the glottal domain features provided better discrimination between the
depressed and control groups compared to prosodic features with an overall classification
accuracy of up to 96%.

Study from Cohn et al., compared the used of vocal prosody with facial image
features and manual facial behaviour annotation in the recognition of depressed and
control groups. Classification accuracy of 79% was reported with the vocal prosody
feature.

An extensive investigation into the contributions and classification of different
types of acoustical features in depressed and non depressed adolescents were compared in
the work of Low et al. Five main acoustic feature categories were investigated according
to the physiological and perceptual components in the human speech production system
which includes the prosodic category, glottal category, spectral category, cepstrum
category and non-linear TEO based category. Unlike other studies reported, the use of
naturalistic speech from family interactions provided a more defined analysis for each
feature category investigated in the recognition of depression symptoms where is more
applicable in a real life environment. A large sample of 139 participants (68 depressed
and 71 controls) speech samples was used for the comparison of the five acoustic features
in the classification of depressed and control subjects. It was reported that the TEO
features clearly outperformed all other features and feature combinations. Depending on
the type of TEO features and the classification method, the classification accuracy ranged from 70% to 97% for male speakers and from 58% to 93% for female speakers.

A comparison for classifying clinical depression using acoustical features of spontaneous speech and read speech was investigated by Alghowinem et al. Some low-level descriptors (LDD) features and functional features which includes, pitch, MFCC, energy, intensity, formants and voice quality features was tested on a clinical dataset of 30 depressed and 30 control subjects with a SVM classifier. It was found that using spontaneous speech was better in classifying clinical depression when compared to using read speech. Results also shown that jitter, shimmer, energy and loudness feature groups were more efficient in classifying depressive speech for both the spontaneous and read speech. An average classification accuracy of up to 78.34% was achieved using the log energy feature measured in spontaneous speech.

The investigation of voice quality features by Scherer et al., for a depression classification was conducted on the available participant-pool of 18 depressed and 18 non-depressed based on the Patient Health Questionnaire depression module (PHQ-9) test conducted. The experiments also tested on post-traumatic stress disorder (PTSD). Four types of acoustic features that were relevant in characterizing voice qualities were compared for the depression recognition task. Three of these features were derived from the glottal source domain which is the normalized amplitude quotient (NAQ), quasi-open quotient (QOQ) and the open quotient approximation using MFCC and trained neural network (QQNN). A peakSlope feature which is an effective correlate of the spectral slope of a signal was also used. An overall system performance of 75% was achieved for the depression classification experiment.
As most reported depression analysis to date has been primarily limited to a single channel analysis, recent studies have started to explore the feasibility of a multimodal analysis utilizing multiple combinations of input channels with different fusion techniques. There have been a number of multimodal approaches being explored for affect and emotion recognition task [18], [72], [152], however, the application on depression related studies are still very limited. Due to the complexity and variability of depression related changes in different individuals, a multimodal analysis would provide a more complete investigation for the correlation of image analysis and speech acoustics in depression.

Recent work by Cummins et al. [25], investigated an audio and visual feature data fusion for the recognition of depression. The audio features were supervectors extracted by the GMM-UBM concept using MFCCs. For the visual features, the space-time interest points (STIP) and Pyramid of Histogram of Gradients (PHOG) features were extracted from facial images. The tested system produced an audio baseline RMSE of 14.12 and video baseline RMSE of 13.61. It was reported that the audio-visual feature fusion for the test set results produced a RMSE of 10.62 which outperformed both the individual audio and video baseline results.

A study by Joshi et al. [59], investigated the audio-video multimodal system for depression detection using 3 types of fusion techniques:

1. Feature fusion: Raw features from different modalities (audio/video inputs) are concatenated to a single feature vector.
2. Score fusion: Different scores from the classifier are combined before the classification decision is made. Score fusion by weighted sum and by a new SVM classifier remodelling on the scores were investigated.

3. Decision fusion: The AND and OR operators are used to fuse decisions from the separate decision outputs from the individual SVM classifier. Decision fusion by learning a new, second-stage SVM classifier was also investigated.

Audio features of fundamental frequency (f0), loudness, intensity and Mel-frequency cepstral coefficients (MFCC) were investigated together with the Space-Time Interest Points (STIP) and Local Binary Patterns Three Orthogonal Planes (LBP-TOP) facial image features using support vector machines (SVM) classifier. First, individual features were tested and a highest classification accuracy of 75% was achieved with the intensity features for the audio related input and 76.7% for the STIP video features. Next, feature combinations were tested for the audio and video features separately. There was a recorded increase in accuracy by 8.3% for the audio combined feature and an increase of 5% for the video combined feature. Lastly, the 3 different types of fusion techniques were tested with different types of audio-visual feature combinations. The reported results showed a highest classification accuracy of up to 91.7% for the feature fusion approach, 88.3% for the score fusion approach and 91.7% for the decision fusion approach.

A similar kind of study was conducted by Alghowinem et al. [3], comparing 3 types of fusion methods (feature fusion, score fusion and decision fusion), however, tested only on speech features (pitch, MFCC, energy, intensity, formants and voice quality) across different types of classifiers (Gaussian Mixture Models - GMM, Support Vector Machines - SVM, Hierarchical Fuzzy Signature - HFS and Multilayer Perceptron
Neural Network - MLP). On average across all groups of acoustic features, the decision fusion method provided the best classification performance of 81.13%.

From these observations, the implementation of a multimodal/multichannel system has been effective in improving classification accuracy from a unimodal/single channel system. Although studies using this implementation are still limited in the depression recognition research area, the promising findings could be beneficial for a more complete analysis of an automatic depression detection system.

2.7 Summary

In order to achieve the primary aim of this study, which is to develop an applicable objective method for early prediction of risk for depression, a review of the past literature relating to depression recognition was conducted. Focus was given to methods related to facial image and speech acoustics. It was found that depression methodologies using facial images and speech were limited to detection, with none being applied for prediction. Depression detection is clearly different that of prediction – in part where detection occurs when depression has developed while prediction is where no symptoms were detectable. As the first onset of depressive symptoms typically occurs during adolescence, this study is focused on depression among adolescents. From understanding these, the study is motivated to test the feasibility of using facial image and speech acoustics as potential predictors of depression and to develop improved methodologies in predicting depression from the best predictor features determined.
Chapter Three:

DATABASE FOR PREDICTION OF DEPRESSION

3.1 Preview

The database used to predict risk for depression was obtained as a result of research collaboration between the RMIT University and the ORYGEN Youth-Health Research Centre in Melbourne, Victoria, Australia. The data base was collected at the ORYGEN research and clinical facilities and professionally annotated by highly trained staff at the Oregon Research Institute, Eugene, OR, U.S.A.

The ORYGEN data contained a collection of video recordings of parent-adolescent conversations collected for the purpose of studying factors associated with the high risk for depression in adolescents. The video recordings were accompanied by clinical diagnostic data assessing the mental health of the adolescent participants.

For the purpose of this study, only the facial images and the corresponding audio tracks representing the adolescent participants complemented with clinical diagnostic data were selected from the whole data base.

The following sections describe details of the ORYGEN data base collection and explain why this database was particularly suitable for the early prediction of risk for depression in adolescents.
3.2 The ORYGEN Y-H Database Collection Procedure

As illustrated in Figure 3.1, the database acquisition was divided into two stages:

1) (T1) Selection of participants and the audio-visual conversational recordings, and
2) (T2) Longitudinal follow-up phase.

**Figure 3.1:** Longitudinal data collection procedure.

3.2.1 Stage T1 of Data Collection

During the participant selection stage in T1, a large sample of 2478 potential participants representing Year Six primary school students (age range: 9–12 years) recruited through schools across metropolitan Melbourne, Australia was subjected to an initial mass screening process. The aim of this initial screening was to select a risk-enriched (increased probability of high risk for depression) population of adolescent participants. Potential candidates underwent full clinical assessment of their mental state conducted in groups of 18-25 people in school classrooms using the Early Adolescent Temperament Questionnaire–Revised (EATQ-R) [36]. Details of the T1 stage of data collection are illustrated in Figure 3.2.
Based on the EATQ-R scores which provided a representation of temperament dimensions for risk factors of emotional and behavioral disorders, a gender balanced sample of 245 risk-enriched adolescents was selected for further participation in the data collection procedures.

The initially selected 245 adolescents, now accompanied by their parent(s) underwent further intensive clinical assessment and selection process.

The first part of the intensive clinical assessment included a screening test based on the Schedule for Affective Disorder and Schizophrenia for School-Ages Children: Epidemiologic Version (K-SADS-E) [103]. This test aimed to determine if any of the 245 adolescent participants meets the Axis I disorders criteria for any past or current major depressive disorder (MDD).

The second part of the intensive assessment included questionnaires for both adolescent and parent(s) to complete. These questionnaires facilitated psychological assessments of depression related issues and symptoms and aimed to make sure that all adolescent participants admitted to further stages of data collection do not suffer from depression and did not experience previous episodes of depression.

Adolescent who were diagnosed with depression or had previous episodes of depression were eliminated from the collection process. Out of the 245 families that participated and passed the first intensive assessment, only 200 agreed to participate in this video recording session. All of these 200 adolescent participants were fluent in English.
The clinical assessments were followed by audio-visual recording sessions during which the selected 200 adolescents and their parent(s) conducted naturalistic conversations with their parent(s).

Two different topics of naturalistic family discussions between parents and children were captured:

1) Event-planning interaction (EPI) and
2) Problem-solving interaction (PSI).

Each interaction was digitally recorded during a separate 20 minutes session. Out of the 200 families, only 191 families with 94 female adolescents and 97 male adolescents fully completed stage T1 of the data collection process.

The video recordings of 191 families were annotated at the Oregon Research Institute by trained coders using the Living-In-Family-Environments (LIFE) coding system [54], [55]. This resulted in second-by-second behavioral codes designed to capture the affective interpersonal behavior of participants during the course of the interaction.

3.2.2 Stage T2 of Data Collection

Note that all the 191 adolescent participants at this stage, T1, were diagnosed with no clinical depression having passed the data collection and selection assessments.

As shown in Figure 3.1, stage T1 was followed by stage T2 consisting of the follow up clinical assessments aiming to determine changes in the mental health state of the same 191 adolescent participants. The time duration between T1 and T2 was about 2.5 years. No audiovisual recordings of participants were made in stage T2.
The assessment included only clinical procedure similar to those used in the screening tests conducted during the first stage (T1) of the data collection process. It included a standard set of questionnaires and K-SADS-E screening procedures.

The results of follow-up sessions conducted 2.5 years after the initial stage T1 showed that 15 participants (6 male & 9 female) out of the 191 total, suffered from the Major Depressive Disorder (MDD) and 3 participants (1 male & 2 female) had other mood disorders (OMD). The remaining adolescent participants showed no symptoms of depression or other mood disorders.

3.3 ORYGEN Y-H Database Recording and Annotation

3.3.1 Audio-visual recordings of family interactions

Recordings of the family interactions were of fundamental importance for this research. The conversational video recordings provided audio and visual samples suitable for the design of speech and facial image analysis and classification procedures aiming to predict risk for depression.

All of the 191 adolescents and their parent(s) participated in two different types of family interaction sessions: event-planning interaction (EPI) and problem-solving interaction (PSI). These parent-child interactions were digitally recorded.

The setup of the recording sessions were organised in laboratory conditions, where noise-free recordings of images and speech representing naturalistic discussions between parents and children could be captured using high quality equipment.

Each interaction was recorded during a separate 20 minute session. During the recording session, family members were seated in a quiet laboratory room a few feet
apart to allow proper distance for capturing undisturbed speech from only one speaker (adolescent and parent). Adolescents and parent(s) were recorded on separate digital audio channels using a wireless microphone placed at the chest level in a way that did not restrict speech. The following recording equipment was used:

a) Digital video recording: Canon XM2 3CCD PAL mini-DV camera system.
   - Two separate cameras were directed at each participant. One camera facing the adolescent and the other camera facing the parent(s).
   - The camera was focused only on the upper body region, targeting the facial section to ensure participants faces remained within the field of view.

b) Audio recording: Wireless Sony ECM-C115 electret condenser microphones.
   - The microphones were fixed to the chest of the child and adult participants.
   - Audio signals were collected through the microphones and were recorded on the left and right channels of each camera, respectively.

Each of the two different conversational topics was discussed for lasted 20 minutes, thus resulting in a total of 40 minutes recorded observational data for each family. A brief description of these two types of family interactions can be given as follows:

a) Event-Planning Interaction (EPI)

   Topics of the EPI were identified based on adolescent and parent responses to the Pleasant Events Checklist (PEC). [76]. The list offered 50 types of different activity topics such as for example “planning for a vacation trip”. Each family group was asked to choose from the PEC list up to 5 topics for discussion.
b) Problem-Solving Interaction (PSI)

Topics of the PSI were identified based on adolescent and parent responses to the Issues Checklist (IC), [108]. The list offered 44 different topics of disagreement between adolescents and parents, such as “adolescent talking back to parents”. Each family group was asked to choose from the IC list up to 5 topics for discussion.

**Figure 3.2:** Participant selection procedure in stage T1.
The order of the interactions was fixed. Both sessions were conducted on the same day, one after another. The EPI discussion was always followed by the PSI. This particular order was chosen due to the expectation, that it would be easier for the participants to switch from a positive state likely to be achieved during the EPI session to a negative state likely to be achieved during the PSI session, than the other way around. In this way, the actual expressed emotions were more likely to result from the current interaction rather than be influenced by emotional states carried out from the previous session.

3.3.2 Observational annotation of family interactions

The video recordings were annotated (labelled) at the Oregon Research Institute by highly trained psychologists (observers) using the Living-In-Family-Environments (LIFE) coding system [54], [55]. This resulted in second-by-second behavioral codes designed to capture the affective interpersonal behavior of participants during the course of the interaction. In total, the annotation script contained 10 different affect (emotional) codes and 27 different content (lexical) codes. A list of these 10 affect codes and 27 content codes can be found in Appendix A- Table A1.

The annotation process was based on verbal and paraverbal aspects of speech, facial expressions, body posture and movement as observed by the markers; no information about felt emotions was collected from participants.

The annotation was conducted by two observers, who were blinded to the diagnostic status and characteristics of the participant. The first observer was completing a full length coding of the entire interaction while the second observer coded approximately 20% of the interaction to validate the first observer’s annotation. The first
and the second observers were randomly assigned to each conversation. The intra- and inter-observer agreement was determined and only the coding results that met the agreement criteria were accepted.

Detailed definitions and marking rules for the affect and the content codes can be found in the LIFE coding instructions [54].

3.4 How the ORYGEN Y-H Data Facilitated Prediction of Depression?

The database suitable for the development and testing of the prediction results was of key importance to this research. The prediction process could be validated due to the fact that the recorded video data was complemented by the results of clinical examinations, first conducted at the time when the recordings were made and then successively repeated during annual follow up sessions.

The ORYGEN data provided facial images and speech recordings of 191 healthy, non-depressed adolescents captured at time T1. It also provided information which of these 191 adolescents developed full blown signs of the Major Depressive Disorder (MDD) 2.5 years later (at time T2). It was therefore possible to assume that the analysis of facial images and voice recordings of adolescents who did, and did not developed depression within the next 2.5 years could be used to validate an image or speech based prediction method for depression.

Since the facial images and audio-recordings were available only for the initial data collection at T1, it was not possible to develop a prediction method guided by changes of facial expressions or acoustic parameters of speech resulting from the development of the full blown symptoms of depression. However, the guidance for the
As described in further sections of this thesis, this assessment at T2 was used to train an automatic image or speech classification procedures that separates the non-depressed image/speech data (collected at T1) of adolescent participants into two groups:

- one including participants who did not develop depression within two years from the time of speech data collection, called the “not at risk” (NAR) group, and
- one including participants who developed depression, called the “at risk” (AR) group.

The AR group consisted of 15 (6 male and 9 female) adolescents, who were non-depressed at T1, but developed major depressive disorder (MDD) within the next 2.5 years (between T1 and T2). To match the sample sizes and the gender ratio, the NAR (control) group was composed of 15 randomly selected adolescents (6 male and 9 female) from a larger pool of participants who were consistently diagnosed as non-depressed and not having other mood disorders in years 1 to 2.5. The gender ratio of 6/9 approximately reflected the well known trend in depression epidemiology, where almost twice as many females as males likely to develop depression during adolescence [26].

3.5 Summary

The ORYGEN Y-H database containing audio-visual recordings of parent-adolescent conversations described in this chapter facilitated research and development of automatic depression prediction systems for an early prediction of risk for depression in
adolescents. This database was particularly necessary in order to test a prediction methodology for depression before symptoms are evident which is in line with the thesis aims for this forerunning investigation into the possibility of predicting clinical depression using image and speech samples. This was different to existing database with similar annotation methods in [54], [55], where, no longitudinal data collection process for depression development studies was done and focused on subjects that were already medically diagnosed with depression. The unique combination of the two stage (T1 & T2) data collection process provided a basis for an efficient validation of speech and image based prediction methodology. Details of these methodologies and the test results are presented in the following chapters. A more detailed description of the psychiatric assessments, data collection process and other implementations of this clinical database can be found in [147], [149].
4.1 Preview

This chapter explores the possibility of using facial images to determine if a non-depressed person is likely to develop clinical depression within the next 1-2.5 years.

The prediction method used to test this possibility followed a typical classification approach of training and testing. Class models were determined using image data from adolescents who were either “at risk” (AR) or “not at risk” (NAR) of depression. As explained in Chapter 3, the actual “risk for depression” necessarily to validate the accuracy of the prediction algorithm was confirmed by the diagnostic outcomes of follow-up clinical examination of participants conducted 2.5 years after audio-visual recordings were made.

Section 4.2 explains the formation of image corpus for training and testing/validation of the prediction algorithm. The image based prediction algorithm is discussed in Section 4.3 which includes the capturing of frontal facial images and image pre-processing, feature extraction techniques used to represent facial image and the image modelling and classification procedure. Section 4.4 presents the depression prediction
results and evaluates the algorithm’s performance. Finally, the summary and conclusions are included in Section 4.5.

4.2 Image Corpus

The prediction method described in this chapter was validated and tested using an image corpus created from the ORYGEN Y-H audio-visual recordings. As described in Chapter 3, all adolescent participants depicted in the collected images were diagnosed as having no current or previous symptoms of depression (at time T1, this is, when these images were captured).

The original recorded image sequences depicted two types of parent-child interactions (EPI and PSI) and each interaction was of approximately 20 minutes duration. The images were sampled at a rate of 25 frames per second.

For the purpose of the prediction algorithm development and testing, image sequences depicting only the fronts of faces of all 191 adolescent participants were extracted from the original recordings.

From these recordings images representing 15 adolescents (9 females and 6 males) who developed full blown signs of the Major Depressive Disorder (MDD) within 2.5 years (at T2) were extracted to create a sub-set representing the “At Risk” (AR) diagnostic class. The remaining pool of image sequences consisting of the 176 adolescents who did not developed signs of depression was used to randomly select a matching sample of 15 adolescents (9 female and 6 male) representing the “Not At Risk” (NAR) diagnostic class.
The female to male ratio reflected the common trend in depression epidemics with almost twice as many females as males likely to develop depression during adolescence [26]. Table 4.1 shows details of the image data for 15 participants from the AR group. The 15 participants from the NAR group were selected at random from the pool of 176 available participants for each cross validation experiment; therefore the NAR information is not included in Table 4.1. However the general distribution of frame numbers for the NAR group was very similar to the AR group.

Note that the number of image frames per session for each participant shown in Table 4.1, are frontal face image sequences that was obtained after image pre-processing.

**Table 4.1: Image Corpus of “At Risk” (AR) Participants**

<table>
<thead>
<tr>
<th>Participant ID</th>
<th>Status</th>
<th>Gender</th>
<th>No. of image frames (EPI, PSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1021</td>
<td>At Risk</td>
<td>Male</td>
<td>8750, 7550</td>
</tr>
<tr>
<td>1915</td>
<td>At Risk</td>
<td>Male</td>
<td>20850, 21775</td>
</tr>
<tr>
<td>2303</td>
<td>At Risk</td>
<td>Male</td>
<td>13100, 20925</td>
</tr>
<tr>
<td>5116</td>
<td>At Risk</td>
<td>Male</td>
<td>11050, 21150</td>
</tr>
<tr>
<td>6113</td>
<td>At Risk</td>
<td>Male</td>
<td>21425, 28775</td>
</tr>
<tr>
<td>8516</td>
<td>At Risk</td>
<td>Male</td>
<td>19275, 12150</td>
</tr>
<tr>
<td>0306</td>
<td>At Risk</td>
<td>Female</td>
<td>9950, 18250</td>
</tr>
<tr>
<td>0313</td>
<td>At Risk</td>
<td>Female</td>
<td>30025, 15750</td>
</tr>
<tr>
<td>0410</td>
<td>At Risk</td>
<td>Female</td>
<td>7300, 7500</td>
</tr>
<tr>
<td>2343</td>
<td>At Risk</td>
<td>Female</td>
<td>9825, 10625</td>
</tr>
<tr>
<td>5008</td>
<td>At Risk</td>
<td>Female</td>
<td>19925, 20750</td>
</tr>
<tr>
<td>5528</td>
<td>At Risk</td>
<td>Female</td>
<td>21225, 22550</td>
</tr>
<tr>
<td>6421</td>
<td>At Risk</td>
<td>Female</td>
<td>18780, 20150</td>
</tr>
<tr>
<td>6910</td>
<td>At Risk</td>
<td>Female</td>
<td>26970, 29850</td>
</tr>
<tr>
<td>9321</td>
<td>At Risk</td>
<td>Female</td>
<td>28725, 30150</td>
</tr>
</tbody>
</table>
4.3 Image Based Depression Prediction Algorithm

4.3.1 Algorithm overview

Figure 4.1 depicts the general framework of the proposed image based prediction methodology.

In general, the prediction procedure consisted of two stages: training and testing. In the first stage (training), characteristic features extracted from the pre-processed images representing known classes were used to train models of the two diagnostic classes AR and NAR. In the second stage (classification or testing), characteristic features extracted from pre-processed images representing unknown classes were “compared” with the models to determine emotional classes (AR or NAR) to which they belonged. The following sections describe the image prediction algorithm with more details.
4.3.2 Image pre-processing

Image sequences for each adolescent representing two separate video sessions, each of which lasting 20 minutes at a capture rate of 25 frames per second were extracted. This resulted in a total of approximately 30000 image frames per person and per video session. The raw image frames were subjected to the following pre-processing steps where unwanted frames can be removed and only quality frames containing useful information to be used in subsequent feature extraction stage.

The pre-processing stage included: image normalization and frontal face detection. These pre-processing processes will first convert all frames to gray scale images which provides gray level information only, then remove frames which do not contain frontal face poses. Only the frontal face image will be cropped out and thus eliminating any background information. The following sub-sections explain in further detail the methodology used for performing each pre-processing task and also explains the advantages of doing so.

4.3.2.1 Image normalization

The raw image frames were first converted to gray scale images which eliminate the hue and saturation information while retaining the luminance or intensity information. The luminance is more important in distinguishing visual features compared to colour images. Using gray scale images also reduces the amount of data per pixel from RGB images which would significantly reduce processing time and complexity.
4.3.2.2 **Frontal face detection using local Successive Mean Quantization Transform (SMQT) features and Split up Sparse Network of Winnows (SNoW) classifier.**

Face detection is an important pre-processing step in facial image related tasks. The advantage of a face detection procedure is that image frames which do not depict frontal face view can be identified and discarded or process to cut out the frontal face area and remove the irrelevant background parts.

Numerous methods of automatic face detection algorithms have been used to capture frontal face images for this purpose. The Viola-Jones detector [144] is among some of the popular face detection algorithms used by researchers. The Viola-Jones detector uses Haar features along with the AdaBoost algorithm, where rectangular regions are compared to categorize sections of an image. This helps to determine if those sections are parts of a facial region which should be kept, or belong to the background, in which cases should be removed. Other frequently used methods include appearance-based models and template matching techniques [24], [69], [116], [131].

In this study, the Successive Mean Quantization Transform (SMQT) features were applied together with the split up SNoW classifier [94] for the frontal face detection. The SMQT [93] method is a fast and effective frontal face detector particularly suitable for the application with our database, where image recordings were made using individual cameras facing individual participants and focused on the upper body region.

The advantage of analysing frontal face poses, in contrast to side views is that they enable a full capture of facial expressions from the entire facial region, where most facial features can be observed. In addition, the use of only frontal poses provides consistent representation of all images being processed.
The SMQT is a transform that reveals the structure of data, or in this case an image. If the input to the SMQT \( D(x) \), is a local area in an image and \( x \) is one pixel with \( L \) being the level of transform, the output from the SMQT transformation, \( M(x) \) will be the extracted illumination insensitive features:

\[
SMQT_L : D(x) \rightarrow M(x)
\]  

(4.1)

The level of transform \( L \), contributes to the number of Mean Quantization Unit (MQU) by a factor of \( 2^{L-1} \). The MQU was calculated in three separate steps:

1) calculation of the mean,

2) quantization and

3) splitting the input set.

Local SMQT features were then classified using the split up Sparse Network of Winnows (SNOW) classifier. A comparison between face and non-face features was made through a defined lookup-table and a decision was made based on a defined threshold.

For the purpose of detecting faces in an image, we used a local area of 3x3 pixels and level \( L=1 \) for the transformation. Local SMQT features were extracted from this local area and classified. The local area was then shifted by a \( \Delta x = 1 \) and \( \Delta y = 1 \) pixels to examine the entire image. A more detailed description of the SMQT face detection method can be found in [93], [94].

4.3.3 Facial image feature extraction

Detected face regions were cropped out from the original image frame of size 1024x1024 pixels to a smaller size of 94x94 pixels. The 2-dimensional image arrays were then
converted into 1-dimensional vectors of size 8836 and used in the process of feature extraction.

Feature extraction methods such as the eigenfaces and fisherfaces have been effectively used for face recognition and other image classification tasks. In this thesis, usefulness of these methods in prediction of clinical depression was investigated.

The eigenface method also known as the principal component analysis (PCA) applied on facial images [138], maps images into lower dimensionality representations and it is commonly used for data dimensionality reduction.

The fisherface method combines the PCA with the linear discriminant analysis (LDA) which uses class specific linear methods for dimensionality reduction [14]. The LDA forms a scatter that is more reliable for classification by maximizing the between-class variance & minimization of within-class variance whereas the PCA maximizes the scatter of all projected samples.

Experiments using these features aims compare facial characteristics differences between AR and NAR groups. This can be modeled by the group of principal values computed by PCA for each category. The different gray values selected based on PCA represented by each image trained for the respective AR and NAR group may characterise a similarity between the group itself and difference between the two groups. For example, a modeled component values for the AR group are more focused on a certain facial region than the NAR group. The following sub-sections discuss the computation of the eigenface (PCA) and fisherface (PCA+LDA) features for the image based prediction of clinical depression in adolescents.
4.3.3.1 Eigenfaces (PCA)

The principal component analysis (PCA) searches for a group of principal values (components) which can represent the original data vector using lower dimensionality. The PCA is therefore frequently used for data reduction purposes.

The application of the PCA to the facial image matrix generates image representations known as eigenfaces (see Figure 4.2).

Principal component analysis (PCA) determines a set of eigenvectors calculated for a covariance matrix of an array of images. These eigenvectors represent basic images also called eigenfaces. A linear combination of eigenfaces can be used to approximate different types of human faces.

PCA helps to reduce the feature dimensionality by mapping the original array of images into a lower-dimensionality space. The application of the eigenfaces (PCA) method followed steps described in [28], [138].

First a training set of face images where each image $I(94 \times 94)$ pixels was defined as a vector of dimension $N = l^2$ was created. A training array $T (N \times M)$ of face images was generated, where each of the $M$ columns represented an $N$-dimensional image vector. A mean face vector of the training array was calculated and subtracted from the original array $T$ and a scatter matrix $S$ was obtained. The mean face of the training image set of $M$ vectors was calculated by the average shown in equation (4.2). The aim was to create a new lower dimensional ($p$-dimensional) subspace of face images known as the face space by performing PCA where the basis vector corresponds to the maximum variance of the original image space. The scatter matrix $S$ defined by equation (4.3) is a set of basic vectors where the PCA of this matrix creates a face space.
\[ \mu = \frac{1}{M} \sum_{i=1}^{M} x_i \]  
\[ S = \sum_{i=1}^{M} (x_i - \mu)(x_i - \mu)^T \]

\(x_i\) is the \(i\)th image with its columns concatenated in a vector.

The eigenvectors (eigenfaces) and the eigenvalues of the array \(S\) were calculated. In order to reduce the data dimensionality only the eigenfaces that represent the largest eigenvalues were kept. This way the original dimension for a single image representation was reduced from \(N=8836\) dimensions to \(p=500\). The training and the testing images were then projected onto the \(p\)-dimensional space which means that each image was represented as linear combinations of the \(p\)-dimensional eigenfaces. The weights obtained

\[x_i \] is the \(i\)th image with its columns concatenated in a vector.

The eigenvectors (eigenfaces) and the eigenvalues of the array \(S\) were calculated. In order to reduce the data dimensionality only the eigenfaces that represent the largest eigenvalues were kept. This way the original dimension for a single image representation was reduced from \(N=8836\) dimensions to \(p=500\). The training and the testing images were then projected onto the \(p\)-dimensional space which means that each image was represented as linear combinations of the \(p\)-dimensional eigenfaces. The weights obtained

\[x_i \] is the \(i\)th image with its columns concatenated in a vector.

The eigenvectors (eigenfaces) and the eigenvalues of the array \(S\) were calculated. In order to reduce the data dimensionality only the eigenfaces that represent the largest eigenvalues were kept. This way the original dimension for a single image representation was reduced from \(N=8836\) dimensions to \(p=500\). The training and the testing images were then projected onto the \(p\)-dimensional space which means that each image was represented as linear combinations of the \(p\)-dimensional eigenfaces. The weights obtained

\[x_i \] is the \(i\)th image with its columns concatenated in a vector.

The eigenvectors (eigenfaces) and the eigenvalues of the array \(S\) were calculated. In order to reduce the data dimensionality only the eigenfaces that represent the largest eigenvalues were kept. This way the original dimension for a single image representation was reduced from \(N=8836\) dimensions to \(p=500\). The training and the testing images were then projected onto the \(p\)-dimensional space which means that each image was represented as linear combinations of the \(p\)-dimensional eigenfaces. The weights obtained
from these linear combinations were used as the training feature vectors stored in two arrays: \( W_{pAR} \) for the AR training class and \( W_{pNAR} \) for the NAR training class.

Figure 4.2 show examples of reduced-dimensionality facial images obtained as an outcome of the PCA process. Note that these images are generated from the FERET (FacE REcognition Techonology) database [106] and do not contain images of any participants from our testing database from ORYGEN due to human ethics agreements.

4.3.3.2 Fisherfaces (PCA+LDA)

The fisherface method [14], [28] combines PCA with the linear discriminant analysis (LDA) which is used to find the vectors in the space that provide best discrimination between the among classes. The PCA component is used as a pre-processing step for the LDA setup. The main concept of LDA is to maximize the between class scatter matrix \( S_B \) and to minimize the within class scatter matrix \( S_W \) which are defined by equation (4.4) for the former and equation (4.5) for the later.

\[
S_B = \sum_{i=1}^{C} M_i (x_i - \mu)(x_i - \mu)^T \tag{4.4}
\]

\[
S_W = \sum_{i=1}^{C} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T \tag{4.5}
\]

\( C \) represents the number of classes where \( M_i \) is the number of training samples in class \( i \), \( \mu_i \) is the mean image of class \( X_i \) and \( x_k \) is the \( k \)-th image of class \( X_i \).

The projection matrix \( W_{LDA} \) is a set of the eigenvectors with the largest eigenvalues of \( S_W^{-1}S_B \). The fisherface method avoids \( S_W \) being singular using the PCA as a pre-processing
step by projecting the image set to a lower dimensional space. Thus, by integrating the LDA with PCA, the optimal projection is defined by equation (4.6).

\[ W_{opt}^T = W_{LDA}^T W_{PCA}^T \]  

(4.6)

In the described experiments, two sets of weights were generated: \( W_{optAR} \) for the AR training class and \( W_{optNAR} \) for the NAR training class.

### 4.3.4 Image modeling and classification

Figure 4.3 shows an overview of the modelling and classification process for the eigenface and fisherface features. Two separate training subsets for the AR and NAR classes were selected for the generation of the lower p-dimensional image feature sets. Two different p-dimensional image sub-spaces were created:

1) A p-dimensional image sub-space representing the AR class. Obtained from the training subset of the AR class.

2) A p-dimensional image sub-space representing the NAR class. Obtained from the training subset of the NAR class.

During the training process, the full training datasets of the respective classes (AR & NAR) were projected onto the corresponding p-dimensional sub-spaces generating sub-sets of lower dimensionality images. The projection onto the image subspace provided the principal weight values that represented the individual classes.

The training weights obtained during this training process included:

1) Training weights from the eigenface feature projected onto the eigenface AR subspace (\( W_{pAR} \))
2) Training weights from the eigenface feature projected onto the eigenface NAR subspace ($W_{pNAR}$)

3) Training weights from the fisherface feature projected onto the fisherface AR subspace $W_{optAR}$

4) Training weights from the fisherface feature projected onto the fisherface NAR subspace $W_{optNAR}$

Similarly, during the testing stage of the experiment, testing weights were obtained by projecting the testing images onto the lower $p$-dimensional space of the AR and NAR classes. For each individual testing image, the following four different testing weights were obtained:

1) $W_{tpAR}$ (projection on the subspace of $W_{pAR}$)

2) $W_{tpNAR}$ (projection on the subspace of $W_{pNAR}$)

3) $W_{toptAR}$ (projection on the subspace of $W_{optAR}$)

4) $W_{toptNAR}$ (projection on the subspace of $W_{optNAR}$).

The feature parameters were classified into the AR and NAR classes using the Nearest Neighbour (NN) classifier [23]. The NN method searches for its nearest class using the standard Euclidean squared vector distance measure given in equation (4.7).

$$d = \|x - y\|^2$$  \hspace{1cm} (4.7)

Where $x$ represented the weight vector of the tested image and $y$ was the weight vector representing a given class. The class (AR or NAR) that provided the smallest value of $d$ was assigned to the input vector.

If the classification result indicated the AR class then the output from the classifier was given as $+1$ and if the result was NAR, then the output was given as $-1$. 
**4.4 Experiments for Image Based Depression Prediction**

**4.4.1 Experimental setup**

The experiments tested a two-class prediction problem in which the facial images were classified into two classes: “At Risk” of depression (AR) and “Not At Risk” (NAR) of depression.

Two types of the predictive classification were tested: person independent and person dependent. Although, the gender dependency has been reported in previous depression detection studies [74], [75], [77], [85], a relatively small number of data representing different genders prohibited these experiments from testing this otherwise important factor.
The assessment of the classification results was based on three parameters: specificity, sensitivity and accuracy typically used in the assessment of diagnostics and binary classification process [10], [75], [85]. These parameters were defined as follows:

\[
\text{Sensitivity} = \frac{TP}{TP+FN} \times 100\% \tag{4.8}
\]

\[
\text{Specificity} = \frac{TN}{TN+FP} \times 100\% \tag{4.9}
\]

\[
\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{4.10}
\]

Where, TP represented the number of true positive outcomes (the number of AR adolescents classified as AR), FP represented the number of false positive outcomes (the number of NAR adolescents classified as AR), TN represented the number of true negative outcomes (the number of NAR adolescents classified as NAR) and FN represented the number of false negative results (the number of AR adolescents classified as NAR).

Two types of tests were carried out to examine the system performances:

1) Person independent test: The number of images classified correctly or incorrectly as AR or NAR was calculated.

2) Person dependent test: The number of adolescents classified correctly or incorrectly as AR or NAR was calculated. An adolescent was then classified as AR if more than 50% of images representing this adolescent were correctly classified as AR. Similarly, an adolescent was classified as NAR more than 50% of utterances representing this adolescent were correctly classified as NAR.
In the person independent as well as the person dependent tests, approximately 50% of the dataset, 7 NAR adolescents and 8 AR adolescents were used for training and the remaining 50% were used for testing. The training and testing processes were repeated for 3 turns of cross validation, and the results were averaged over these three runs. This process was conducted independently for three types of family interaction combinations used during the recording sessions (EPI, PSI and combined EPI+PSI) and for two types of feature extraction/data compression methods (eigenface/PCA and fisherface/PCA+LDA). The system performance was determined using the parameters given in equations (4.8)-(4.10).

### 4.4.2 Person independent classification results

In the case of a person independent prediction, all available images for the 15 AR and 15 NAR adolescents were put together (i.e. the data was not separated for each adolescent) and used to generate the class balanced training and testing sets for the classification process.

Table 4.2 shows the performance results for the person independent classification using two different feature extraction methods and three different types of family interactions. The percentage of AR images classified correctly as AR class is shown in the sensitivity column and the percentage of NAR images correctly classified as NAR class is shown in the specificity column. The overall classification performance in correctly classifying AR and NAR images is found in the accuracy column of Table 4.2.

In general the overall classification accuracy obtained from both feature set and across all different family interactions were less desirable for the person independent
classification approach. The highest classification accuracy came from the fisherface features experimenting on the PSI interaction session, was only able to provide a percentage of 51%, which is just a chance level prediction rate. The sensitivity performance of the fisherface features resulted in an improvement by 10%, 6% and 9% for the EPI, PSI and EPI+PSI sessions respectively compared to the eigenface features.

Table 4.2 PERSON INDEPENDENT CLASSIFICATION PERFORMANCE RESULT

<table>
<thead>
<tr>
<th>Training/Testing Features</th>
<th>Training/Testing Dataset</th>
<th>Classification Accuracy in %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
</tr>
<tr>
<td>Eigenface (PCA)</td>
<td>EPI session only</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>PSI session only</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>EPI+PSI session</td>
<td>46</td>
</tr>
<tr>
<td>Fisherface (PCA+LDA)</td>
<td>EPI session only</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>PSI session only</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>EPI+PSI session</td>
<td>55</td>
</tr>
</tbody>
</table>

There was a slight decline in the classification percentage of the specificity performance for the fisherface features when compared with the eigenface features in the EPI and PSI sessions of 4% and a slight increase of 5% in the EPI+PSI session. However, the overall system classification accuracy was higher for the fisherface features across all interaction sessions compared to the eigenface features.

Two distinct observations can be made from the person independent classification results:

1) The PSI interaction session performed the best among all other interaction combinations for both the eigenface and fisherface feature in terms of overall system classification accuracy.
2) The fisherface feature performed best across all interaction sessions when compared to the eigenface feature in terms of overall system classification accuracy.

4.4.3 Person dependent classification results
A person based classification is important for a medical diagnosis analysis in determining if an adolescent is “at risk” of clinical depression or clear from the risk of clinical depression. In order to perform such classification task, a reasonable threshold has to be selected for the number of classified images as “at risk” (AR) or “not at risk” (NAR) coming from an adolescent. For instance, if the number of AR classified images for an adolescent was above a certain selected threshold, then the adolescent will be classified as an AR adolescent. Similarly, if the number of NAR classified images for an adolescent was above a certain selected threshold, then the adolescent will be classified as an NAR adolescent.

Section 4.4.3.1 explains the procedure of selecting the optimal threshold for the person dependent classification decision making and Section 4.4.3.2 reviews the results obtained from the experiment conducted using this approach.

4.4.3.1 Selection of an optimal decision rule for the person dependent classification
From the medical prevention point of view, a predictive diagnostic method for depression should have a relatively large true positive rate (TPR) i.e. sensitivity, at the cost of lower values of false positive rate (FPR) i.e. 1-specificity. TPR to FPR ratio was controlled by the classification decision rule based on two thresholds. The first threshold \( \zeta_{\text{AR}} \) specifying
the percentage of images classified as AR above which the overall classification for a given AR adolescent was determined as AR, and the second threshold, $\zeta_{\text{NAR}}$ specifying the percentage of images classified as NAR above which the overall classification for a given NAR adolescent was determined as NAR. The classification results based on each of the 9 decision rules listed in Table 4.3 were used to calculate the parameters given in Equations (4.8)-(4.10) respectively.

The first three sets of 9 points corresponding to the 3 interactions (EPI, PSI and EPI+PSI) and the fisherface features were plotted in Figure 4.4 (a), whereas the three sets of 9 points corresponding to the 3 interactions and the eigenface features are plotted in Figure 4.4 (b). The ROC (Receiver Operating Characteristic) plots in Figure 4.4 were then used to determine which of the 9 classification decision rules provides the best TPR to FPR ratio. The point closest to the upper left corner of the ROC plane represents the best decision rule. It was therefore determined that best combination giving approximately the highest TPR while having a reasonable un-skewed FPR (i.e. not having a very high TPR score in the expense of very low FPR) was a balanced 50% criterion corresponding to classification decision rule number 5 in Table 4.3.
Table 4.3 CLASSIFICATION DECISION RULES

<table>
<thead>
<tr>
<th>Rule</th>
<th>ζAR</th>
<th>ζNAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10%</td>
<td>90%</td>
</tr>
<tr>
<td>2</td>
<td>20%</td>
<td>80%</td>
</tr>
<tr>
<td>3</td>
<td>30%</td>
<td>70%</td>
</tr>
<tr>
<td>4</td>
<td>40%</td>
<td>60%</td>
</tr>
<tr>
<td>5</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>6</td>
<td>60%</td>
<td>40%</td>
</tr>
<tr>
<td>7</td>
<td>70%</td>
<td>30%</td>
</tr>
<tr>
<td>8</td>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>9</td>
<td>90%</td>
<td>10%</td>
</tr>
</tbody>
</table>

4.4.3.2 Results of the person dependent classification using the optimal classification decision rule

Table 4.4 shows the results of person dependent classification accuracy based on the 50% criterion determined in Section 4.4.3.1. Similar to the two distinctive observations from the results shown in the person independent classification in Table 4.2, results in Table 4.4 also confirms that the fisherface feature and the PSI session perform consistently better than the rest of the experiments. This is consistent with previous studies showing that the LDA performed better than PCA in various cases of face recognition [14], [117].

In this experiment, the fisherface method performed slightly better than the eigenface method. Also the PSI session was consistently giving the highest accuracy. This finding is consistent with other clinical depression experiments conducted by Low et al. [74] and has been supported by the fact that the PSI session involves conflictual
interactions which are strong correlates of adolescents’ depression in family environments where often negative emotions are expressed [128].

In terms of sensitivity classification percentage, the fisherface features improved by an average of 16% across all interaction sessions compared to the eigenface features. Although the average specificity classification percentage across all interaction sessions for the fisherface features was lower by 9% to the eigenface features, it was more desirable to have a higher sensitivity ratio at the cost of lower specificity (i.e. more correct classification of AR adolescents than more correct classification for NAR adolescents) in this experiment.

![ROC Space (Fisherface)](image1) ![ROC Space (Eigenface)](image2)

(a) Fisherface feature ROC  (b) Eigenface feature ROC

**Figure 4.4 ROC space analysis**

Across all interaction sessions, the fisherface features were also performing consistently better in its overall system classification accuracy by and increase if 7% and 4% in the EPI session and PSI session respectively, and the same classification accuracy
of 52% was obtained in the EPI+PSI interaction session combination for both types of features.

Table 4.4 PERSON DEPENDENT CLASSIFICATION PERFORMANCE RESULT

<table>
<thead>
<tr>
<th>Training/Testing Features</th>
<th>Training/Testing Dataset</th>
<th>Classification Accuracy in %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
</tr>
<tr>
<td><strong>Eigenface (PCA)</strong></td>
<td>EPI session only</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>PSI session only</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>EPI+PSI session</td>
<td>38</td>
</tr>
<tr>
<td><strong>Fisherface (PCA+LDA)</strong></td>
<td>EPI session only</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>PSI session only</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>EPI+PSI session</td>
<td>60</td>
</tr>
</tbody>
</table>

4.5 Summary and Conclusions

The search for an efficient prediction methodology for depression started with image based techniques. This chapter investigated the feasibility of using facial features in an automatic prediction of clinical depression in adolescents.

The Prediction task was performed using a classification approach based on two class models trained on a long-term observational data and representing adolescents that are “at risk” (AR) of developing depression within 1-2.5 years and adolescents that are “not at risk” (NAR) of depression. In the classification process, the eigenface (PCA) feature extraction and data reduction algorithm was compared with the fisherface (PCA+LDA) method. In both cases the nearest neighbour (NN) classifier was used. Image samples used in the experiments were recorded during two types of family interactions: PSI (problem solving interaction) and EPI (event planning interaction).
The results showed that the highest prediction accuracy was achieved using the fisherfaces (PCA+LDA), providing 51% accuracy for the person independent classification during the PSI session and 61% when using the person dependent approach.

These results are relatively low, and do not provide a definite answer to the question of whether it is possible to predict depression from facial images before a conventional diagnosis can be made. However, it is interesting that the PSI session provided consistently higher prediction results than the EPI session and the combination of PSI and EPI sessions. In the case of PSI, it was possible to obtain classification accuracy that was about 7%-11% above a pure chance level (50%). This can be attributed to the fact that during the problem solving task a controversy is more likely to be elicited between speakers than during the event planning session. This in turn could lead to an increased effort to control the behavior and to present “nice” or “neutral” facial expressions during the recording sessions. The higher classification rates at the PSI session could therefore indicate that there is a difference in the degree to which the AR and NAR adolescents can control their facial expressions.

In conclusion, results from the image based experiments did not provide strong support for the use of facial images as predictors of depression. Further research using different types of facial image based features and classification techniques needs to be carried out in order to validate the feasibility of predicting depression from facial images. This however was a preliminary experiment carried out in order to test the automatic prediction of clinical depression using signal processing methodology as proposed in this thesis.
These conclusions led to the investigation of vocal features from adolescent’s voice recordings as hopefully better predictors of depression. This topic is discussed in the next chapter.
Chapter Five:

SPEECH BASED PREDICTION OF DEPRESSION

5.1 Preview

This chapter explores the possibility of using speech signals to determine if a non-depressed person is likely to develop clinical depression within the next 1-2.5 years. The prediction method used to test this possibility is based on classification of feature parameters derived from speech. The chapter starts with an investigation of classical, single-channel approaches, where the modelling process and the classification decision procedures are based on a single type of features (Glottal, Prosodic, TEO or Spectral) and proceeds into a new, advanced multi-channel system with each channel performing independent classification based on a different type of features.

The primarily questions asked here are:

- Is it possible to predict the risk for depression from acoustic speech parameters 2.5 years before the full blown symptoms occur?
- What kind of acoustic parameters provide the best discrimination between the “At Risk” (AR) and “Not AT RISK” (NAR) adolescents?
- What is the most efficient speech modeling and classification procedure for an early prediction of depression?
Figure 5.1 illustrates a general framework for the proposed speech based prediction of depression. The main stages of this system include:

- pre-processing and feature extraction,
- single-channel classification and
- multi-channel classification (with or without optimization).

Speech samples for testing and evaluation of the prediction algorithms were selected from the audio track of the ORYGEN Y-H audiovisual recordings. Formation of the speech corpus is described in Section 5.2.

In both, the training and the testing phases of the single-channel classification systems, speech utterances were pre-processed to remove the non-voiced parts and concatenate high quality voiced speech segments. These high quality voiced speech segments were manually screened to remove any voiced parts which contained high noise levels and unclear speech utterances. The pre-processing details are presented in Section 5.3. These voiced speech samples were then processed to calculate feature parameters. The methodology used to derive different types of feature parameters is described in Section 5.4.

During the training phase, the individual feature parameters within each channel were used to generate class specific models (AR and NAR) using the Gaussian Mixture Model (GMM). Whereas during the testing phase, the Bayesian decision procedure was applied to determine the highest likelihood class (AR or NAR) for a given test sample of speech. Details on the modelling and classification procedures applied within a single channel are explained in Section 5.5 and Section 5.7.
The significance of each type of features (Glottal, Prosodic, TEO and Spectral) was statistically assessed to determine parameters that are significant in distinguishing between speech of adolescents who are at risk of depression and those who are not. The methodology and results are described in Section 5.6.

The classification results (depression prediction) obtained from individual single channels was used to compare prediction performance of different types of features. These experiments and results are described in Section 5.8.

When the outcomes of the parallel single-channel classifiers were combined to provide the final classification results, the system acquired a multi-channel character. The performance of the multi-channel structure was tested in order to determine the most efficient way of making the final prediction decision based on outcomes provided by individual channels. These experiments and results are described in Sections 5.9 to 5.11.

When investigating the proposed multi-channel system, the focus was on finding the most efficient way of combining the weighted outputs of individual channels into the final prediction output. Investigations of different methodologies for determining the weight parameters and the corresponding experimental results are described in Sections 5.10 and 5.11.

Finally, the summary and conclusions of the speech based prediction of depression can be found in Section 5.12.
5.2 Speech Corpus

The speech data used to train and test/evaluate the speech based depression prediction algorithms was formed from a sound track of the audio-visual recordings of parent-child collected by the ORYGEN-Youth Health Research Centre in Melbourne, Australia.

The audio signals were collected via wireless Sony ECM-C115 electret condenser microphones fixed to the chest of the child and adult participants, in a way that did not restrict the person’s speech. Separate channels were used for the adolescent and parents voice recordings. For the purpose of this study, only one channel (left or right) containing the adolescent’s voice recording was used. The sampling rate for the audio track was 44 kHz.

As previously described in Chapter 3, the ORYGEN Y-H data collection process was conducted in two stages. During the first stage (T1) a group of 191 (94 female & 97 male) adolescents (12-13 years of age) was selected from a larger community sample.
using the standard depression screening criteria. All adolescents within the selected sample were diagnosed as having no symptoms of depression. Apart of clinical examinations, during the first stage, audio-visual recordings of parent-child conversations were made. Two different topics of naturalistic family discussions between parents and children were captured; event-planning interaction (EPI) and problem-solving interaction (PSI). Each interaction was digitally recorded during a separate 20 minutes session. The same group of adolescent participants underwent subsequent follow up diagnostic assessments 2.5 years later (T2) (without audio-visual recordings) to determine which of the adolescents developed depression or other type of mental health problems. The T2 examination showed that 15 participants (6 male & 9 female) developed full blown symptoms of the Major Depressive Disorder (MDD) and 3 participants (1 male & 2 female) were diagnosed with other mood disorders. The remaining adolescent participants had no symptoms of depression or other mood disorders.

More detailed description of the ORYGEN-YH data collection procedures and the psychiatric assessments tasks can be found in Chapter 3.

Since the audio-recordings were available only for the initial data collection at time T1, it was not possible to develop a prediction method guided by changes of acoustic parameters of speech resulting from the development of the full blown symptoms of depression. However, the guidance for the speech classification process could be provided by the diagnostic assessments at time T2 (2.5 years after T1). This assessment was used to train an automatic speech classification procedure that separates the non-depressed speech (collected at T1) of adolescent participants into two groups; one including participants who did not developed depression within two years from the
time of speech data collection, called the “not at risk” (NAR) group, and one including participants who developed depression, called the “at risk” (AR) group.

The AR group consisted of 15 (6 male and 9 female) adolescents, who were non-depressed at T1, but developed major depressive disorder (MDD) between T1 and T2. To match the sample sizes and the gender ratio, the NAR (control) group was composed of 15 randomly selected adolescents (6 male and 9 female) from a larger pool of participants who were consistently diagnosed and non-depressed and not having other mood disorders at time T2. The gender ratio of 6/9 approximately reflected the well known trend in depression epidemics, where almost twice as many females as males likely to develop depression during adolescence.

For each of the 191 adolescent participants a number of speech samples were manually selected for processing by listening to the recordings and cutting out silence and parts that contained overlapping voices (cross-talk) and other types of undesirable noise. After this process, each participant was on average represented by 110 speech samples with each speech sample of an average duration 5-6 seconds.

Table 5.1 shows details of the speech samples for 15 participants from the AR group. The 15 participants from the NAR group were selected at random from the pool of 176 available participants for each cross validation experiment; therefore the NAR information is not included in Table 5.1. However the general distribution of frame numbers for the NAR group was very similar to the AR group.
Table 5.1: Speech Corpus of “At Risk” (AR) Participants

<table>
<thead>
<tr>
<th>Participant ID</th>
<th>Status</th>
<th>Gender</th>
<th>No. of speech samples for PSI session</th>
</tr>
</thead>
<tbody>
<tr>
<td>1021</td>
<td>At Risk</td>
<td>Male</td>
<td>40</td>
</tr>
<tr>
<td>1915</td>
<td>At Risk</td>
<td>Male</td>
<td>43</td>
</tr>
<tr>
<td>2303</td>
<td>At Risk</td>
<td>Male</td>
<td>128</td>
</tr>
<tr>
<td>5116</td>
<td>At Risk</td>
<td>Male</td>
<td>122</td>
</tr>
<tr>
<td>6113</td>
<td>At Risk</td>
<td>Male</td>
<td>65</td>
</tr>
<tr>
<td>8516</td>
<td>At Risk</td>
<td>Male</td>
<td>76</td>
</tr>
<tr>
<td>0306</td>
<td>At Risk</td>
<td>Female</td>
<td>98</td>
</tr>
<tr>
<td>0313</td>
<td>At Risk</td>
<td>Female</td>
<td>97</td>
</tr>
<tr>
<td>0410</td>
<td>At Risk</td>
<td>Female</td>
<td>82</td>
</tr>
<tr>
<td>8516</td>
<td>At Risk</td>
<td>Female</td>
<td>49</td>
</tr>
<tr>
<td>0306</td>
<td>At Risk</td>
<td>Female</td>
<td>151</td>
</tr>
<tr>
<td>0313</td>
<td>At Risk</td>
<td>Female</td>
<td>140</td>
</tr>
<tr>
<td>0410</td>
<td>At Risk</td>
<td>Female</td>
<td>104</td>
</tr>
</tbody>
</table>

5.3 Speech Pre-processing

The aim of the pre-processing stage was to detect and discard unvoiced parts of speech samples and keep only the voiced segments.

The voiced speech segments were detected using the linear prediction (LP) technique for a voice activity detector described in Childers, “Speech processing and synthesis toolboxes” [19].

In order to reduce the volume of data used in the processing, the speech samples were down sampled from the original recording rate of 44 kHz to 11 kHz.

The speech signals were then normalised based on the maximum amplitude and segmented into the hamming window frames of length 25 milliseconds (275 samples) and 50% overlap between frames. In the event where the final frame was less than the frame length, random noise was appended to that frame. In order to remove any low frequency drifts, speech frames were filtered using a zero-phase forward and reverse digital filtering [102], as given in Equation (5.1).
\[ y(i) = b(1) \cdot x(n) + b(2) \cdot x(n - 1) + \ldots + b(nb + 1) \cdot x(n - nb) \]
\[ - a(2) \cdot y(n - 1) - \ldots - a(na + 1) \cdot y(n - na) \]  
(5.1)

Where \( x(n) \) denote the filter input data, \( y(n) \) denote the filtered output data, \( a \), represent the filter recursive coefficients and \( b \), is the filter non-recursive coefficients.

The speech frames were normalized again at a maximum magnitude of 1000 to eliminate any cross-talk voice from the parent’s microphone.

A fixed-frame linear prediction analysis was then performed on the frame-by-frame basis using the 13th order linear prediction coefficients (LPC). The LP analysis adopted the orthogonal covariance method where the energy of the prediction error and the first reflection coefficient, \( r_1 \), were calculated for each frame.

Equation (5.2) shows the calculation of the first reflection coefficient \( r_1 \).

\[ r_1 = \frac{1}{N} \frac{\sum_{n=1}^{N-1} x(n) x(n + 1)}{\sum_{n=1}^{N} x(n) x(n)} \]  
(5.2)

Where, \( x(n) \) is the speech sample of a frame and \( N \) is the number of samples in the frame.

The energy of the signal in an analyzed frame was calculated as the geometric mean of the energy terms for two individual sub frames using Equation (5.3).

\[ E = \sqrt{E_1 \cdot E_2} \]  
(5.3)

Where, \( E_1 = x_1(n)^2 \), \( E_2 = x_2(n)^2 \). The first half of the sub frame is denoted as \( x_1(n) \) and the second half of the sub frame is denoted as \( x_2(n) \).

An optimal threshold value of 0.2 for the first reflection coefficient and 1.85*10^7 for the signal energy was determined experimentally [56], in order to classify
a speech frame as voiced or unvoiced. Thus for a voiced frame the condition was that \( r_i > 0.2 \) and \( E > 1.85 \times 10^7 \) otherwise, the speech frame considered to be unvoiced.

During this pre-processing stage, frames that did not contain voiced speech were discarded, whereas frames containing voiced speech were concatenated and used in the subsequent feature extraction process.

5.4 Calculation of Acoustic Features

This section describes methods used to calculate acoustic speech parameters used in the speech based prediction of depression. The feature parameters were calculated on the frame-by-frame basis.

Following suggestions from previous works on depression classification in speech [39], [75], [84], [85], [104], the following categories of acoustic features were tested:

1) Glottal features (G); derived from the glottal waveform,

2) Prosodic features (P),
   a. Prosodic parameters derived from the speech waveform (PS)
   b. Prosodic parameters derived from the glottal waveform (PG)

3) Teager energy operator (TEO) based features
   a. TEO parameters derived from the speech waveform (TEOS)
   b. TEO parameters derived from the glottal waveform (TEOG)

4) Spectral features (S); derived from the speech waveform.

Table 5.2 contains more detailed summary of these categories, sub-categories and the numbers of parameters falling into each category/sub-category.
### Table 5.2: Summary of Acoustic Feature Categories

<table>
<thead>
<tr>
<th>Main Category</th>
<th>Sub-category</th>
<th>No. Of feature coefficients per frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glottal (G)</td>
<td>Glottal Timing (G-T)</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Glottal Frequency (G-F)</td>
<td>3</td>
</tr>
<tr>
<td>Prosodic derived from speech wave (PS)</td>
<td>Jitter (PS-J)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Shimmer (PS-S)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>F0 (PS-F0)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>LogE (PS-LogE)</td>
<td>1</td>
</tr>
<tr>
<td>Prosodic from derived glottal wave (PG)</td>
<td>Jitter (PG-J)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Shimmer (PG-S)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>F0 (PG-F0)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>LogE (PG-LogE)</td>
<td>1</td>
</tr>
<tr>
<td>TEO features derived from speech wave (TEOS)</td>
<td>TEO-SB-Auto-Env</td>
<td>15</td>
</tr>
<tr>
<td>TEO features derived from glottal wave (TEOG)</td>
<td>Sub-Band 1 (TEOG-SB1)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Sub-Band 2 (TEOG-SB2)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Sub-Band 3 (TEOG-SB3)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Sub-Band 4 (TEOG-SB4)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Sub-Band 5 (TEOG-SB5)</td>
<td>3</td>
</tr>
<tr>
<td>Spectral features derived from speech wave (SS)</td>
<td>Flux</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Centroid</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Entropy</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Roll-off</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>PSD</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>FMTS &amp; FBWS</td>
<td>6</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>69</strong></td>
</tr>
</tbody>
</table>

The last column in Table 5.2 shows the number of feature coefficients from each sub-category used to form the feature vectors for each of the main categories. For example, for classification based on the glottal features (G), the feature vectors contained 12 coefficients; the first 9 coefficients represented values of the glottal time-domain (GLT) parameters and the last 3 coefficients represented values of the glottal frequency-domain (GLF) parameters.
Selection of this wide range of acoustic speech parameters provided the additional benefit of representing speech characteristics that can be related to different psycho-physiological processed and physical mechanics of speech formation. Moreover, the knowledge of individual contributions of these features to the depression-related changes in speech characteristics could advance our understanding of how the very early symptoms of depression are developed. The following sub-sections explain in detail the procedures used to calculate individual feature categories and sub-categories listed in Table 5.2.

5.4.1 Glottal features (G):

The glottal flow waveforms is closely linked to the voice production and thus has many applications to speech related studies. Recent findings have shown that glottal features carries distinctive information in the analysis of speech related to clinical depression [75], [84], [85]. In, [85] the glottal features were shown to provide depression recognition accuracy for depressed adults of up to 87% for males and 94% for females. Similarly high depression recognition rates for glottal features were reported in [75] (up to 74% accuracy for male and female depressed adolescents).

The traditional understanding of how voiced speech comes from the source-filter (SF) model was proposed by Fant [37]. It assumes that the air flow through the vocal folds and vocal tract has a unidirectional and laminar character. The vibration of vocal folds generates a periodic signal known as the glottal wave. The glottal wave is then passed through the vocal tract cavities that act as a filter. The glottal wave modulated by the vocal tract filter configuration with resonant frequencies forms the speech signal. The
Fant’s model given in this form provides basis for calculation of the glottal wave from the speech signal simply, by passing the speech signal through the inverse of the vocal tract filter.

As the human voice production system operates by the air flow through the vocal folds and the vocal tract, and the glottal flow estimate is formed when air flows through the vocal folds, by inversing the estimated vocal tract and lip radiation filters from the source of the speech signal, the glottal flow can be estimated. One common method used is the glottal inverse filtering method which deduces the estimate of the air flow through the glottis from acoustic speech [5], [6], [8], [145]. This is done by canceling the effect of the vocal tract using an inverse model of the tract from an estimated vocal tract filter resulting in a glottal flow estimate.

An iterative adaptive inverse filtering algorithm (IAIF) [5], [8], based on the discrete all-pole modeling (DAP) was used to generate the glottal wave. Using IAIF, the glottal source can be extracted by applying the inverse vocal tract filter with the assumption that the glottal production can be estimated by an all-pole filter. Initially a 1st-order DAP analysis was performed to model the vocal tract and estimate the glottal flow on the speech signal, then a higher-order DAP analysis was carried out to obtain a more accurate estimate of the glottal flow.

For each cycle of the glottal waveform, glottal timing (G-T) and glottal frequency (G-F) features were estimated based on the parameterization of the opening and closing phases of the vocal folds. These features were then averaged over the time duration of a frame.
The glottal timing (G-T) was represented by 9 parameters representing timing of the opening and closing phases of the glottal cycle. These parameters included: open quotients (primary and secondary), open quotient approximation, quasi-open quotient, speed quotients (primary and secondary), closing quotient, amplitude quotient and normalized amplitude quotient. The glottal frequency (G-F) was represented by 3 different frequency domain parameters calculated from the glottal waveform spectrum. These parameters included: the value of the parabolic spectral parameter, amplitudes of the first and the second harmonic components and the harmonic richness factor. The glottal parameters were calculated using the TTK Aparat toolbox [2].

5.4.1.1 Time domain glottal parameters (G-T)

The glottal timing (G-T) was represented by 9 different parameters. These parameters were calculated either directly from the glottal wave or from its first derivative. Parameters derived from the glottal wave were based on the estimates of the opening and closing time intervals, whereas parameters derived from the first derivative used the estimates of minimum and maximum values within each glottal period. Figure 5.2 (a) shows an example of a speech waveform. Measurements used in the calculation of the 9 time domain parameters listed below were extracted from the glottal flow estimate in Figure 5.2 (b) and the glottal flow derivative in Figure 5.2 (c).

Using the parameters marked in Figure 5.2, the 9 G-T feature parameters were estimated as follows.
1. Primary open quotient (OPQ₁)

The primary open quotient is calculated as a ratio of the time duration of the primary opening phase $T_{op1} + T_{cl}$ to the time duration of the glottal cycle $T$:

$$OPQ_1 = \frac{T_{op1} + T_{cl}}{T}$$  \hspace{1cm} (5.4)

2. Secondary open quotient (OPQ₂)

The secondary open quotient is calculated as a ratio of the time duration of the secondary opening phase $T_{op2} + T_{cl}$ to the time duration of the glottal cycle $T$:

$$OPQ_2 = \frac{T_{op2} + T_{cl}}{T}$$  \hspace{1cm} (5.5)

3. Open quotient approximation (OPQₐ)

The approximation of the Open Quotient is derived from the Liljencrants–Fant (LF) model:

$$OPQ_\alpha = A_{pk}(\frac{\pi}{2A_{dmax}} + \frac{1}{A_{dmin}})f_0$$  \hspace{1cm} (5.6)

Where $A_{pk}$ denotes the peak-to-peak amplitude of the glottal wave, $A_{dmax}$ and $A_{dmin}$ are the maximum and minimum amplitudes of the glottal wave derivative respectively and $f_0$ denotes the fundamental frequency of the glottal wave.

4. Quasi-open quotient (QOPQ)

Quasi-open quotient calculates the ratio of the opening phase duration $T_{q50}$ corresponding to 50% of the peak-to-peak amplitude of the glottal wave, $A_{pk}$ to one glottal cycle duration $T$: 
\[ QOPQ = \frac{T_{q50}}{T} \]  

(5.7)

5. Primary speed quotient (\(SQ_1\))

The primary speed quotient is the ratio of the primary opening time interval \(T_{op1}\) to the closing time interval \(T_{cl}\):

\[ SQ_1 = \frac{T_{op1}}{T_{cl}} \]  

(5.8)

6. Secondary speed quotient (\(SQ_2\))

The secondary speed quotient is the ratio of the secondary opening time interval \(T_{op2}\) to the closing time interval \(T_{cl}\):

\[ SQ_2 = \frac{T_{op2}}{T_{cl}} \]  

(5.9)

7. Closing quotient (CLQ)

The closing quotient is the ratio of the closing time interval \(T_{cl}\) to the duration of one glottal cycle \(T\), denoted as:

\[ CLQ = \frac{T_{cl}}{T} \]  

(5.10)

8. Amplitude quotient (AQ)

The amplitude quotient is given as the ratio between the peak-to-peak amplitude \(A_{pk}\) of the glottal wave estimate and the minimum amplitude of the glottal wave derivative \(A_{dmin}\):
9. Normalized amplitude quotient (NAQ)

The normalized amplitude quotient is given as the ratio of the amplitude quotient \( AQ \) to the duration of one glottal cycle \( T \):

\[
NAQ = \frac{AQ}{T}
\]
Figure 5.2: Examples of a speech frame (a), glottal wave (b), first derivative of the glottal wave (c). The parameters marked on these pictures were used to derive the 9 G-T features.

5.4.1.2 Frequency domain glottal parameters (G-F)

The glottal frequency (G-F) features were represented by 3 different frequency domain parameters calculated from the glottal wave spectrum. The glottal spectrum was generated by applying the fast Fourier transform (FFT) to the glottal wave estimate. Figure 5.3 shows an example of the glottal wave spectrum and measurements used in the calculation of the frequency domain parameters listed below.
**Figure 5.3:** An example of the glottal wave spectrum. H1 and H2 denote the magnitude of the first and second harmonics respectively.

The 3 G-F feature parameters were estimated as follows.

1. **Parabolic spectral parameter (PSP)**

   The PSP is given as the ratio of the normalized spectral decay $A$ to the maximum spectral decay $A_{\text{max}}$ of the glottal wave. The spectral decay is computed by fitting the second-order polynomial to the estimated glottal wave spectrum for a single glottal cycle. The detailed procedure for calculating PSP can be found in [7].

   $$PSP = \frac{A}{A_{\text{max}}}$$  \hspace{1cm} (5.13)

2. **First and second harmonic ($H12$)**

   Measures the difference between the formants of the first and second harmonic.

   $$H12 = H1 - H2$$  \hspace{1cm} (5.14)
3. Harmonic richness factor (HRF)

The harmonic richness factor is given as the ratio of the sum of magnitudes of higher harmonics (higher than 2) to the magnitude of the first harmonic.

\[ HRF = \frac{\sum H_{k>2}}{H_1} \]  

(5.15)

A more detailed description of the glottal parameters and calculations used in this study can be found in [6], [75].

5.4.2 Prosodic features derived from the speech wave (PS)

Unlike the glottal and TEO features that relate to speech production, the prosodic features are strongly correlated with speech perception and subjective measures of speech quality such as pitch and loudness. Prosodic measures have been used as indicators of person’s behavior, stress, emotion and more recently in speech based depression diagnosis [39], [75], [85], [104]. In [75], depression recognition using prosodic parameters were reported to yield moderate results of about 59% of accuracy for male and 67% for female depressed adolescents. Since the depressed speech is often characterized as “flat”, “lifeless” and monotonous [86], subtle cycle-to-cycle changes in pitch and loudness captured indirectly by the prosodic features would provide important cues facilitating early prediction of depression in speakers.

The prosodic features were estimated from the speech waveform (PS) and included: fundamental frequency (PS-F0), log of energy (PS-LogE), jitter (PS-J) and shimmer (PS-S).
5.4.2.1 Fundamental frequency (PS-F0)

The PS-F0 parameter represented the fundamental frequency of the vocal fold vibration derived from the speech waveform. It was estimated for each voiced frame of the speech signal using the modified autocorrelation method described in [17]. The algorithm estimated the F0 values in the lag domain which has been proven to be more reliable than the frequency domain F0 estimation methods. Using the lag domain autocorrelation estimate as given in equation (5.16), the pitch periods were found by searching for the maximum point of the autocorrelation function for each speech signal within the recommended range of 40Hz to 1000Hz shown in equation (5.17).

\[
acorr(lag) = \sum_{n=0}^{N-1-lag} x(n)x(n+\text{lag}) \quad (5.16)
\]

\[
F0 = \max(acorr(lag) \ (40:1000)) \quad (5.17)
\]

5.4.2.2 Logarithmic Energy (PS-LogE)

The logE parameter was given as log of the squared amplitudes of speech time waveform samples \(x(n)\).

\[
E_{\text{log}}(m) = \log \sum_{n=1}^{m} x^2(n) \quad (5.18)
\]

5.4.2.3 Jitter (PS-J)

The jitter represented the variation of F0 from cycle-to-cycle. It was estimated as the first derivative of the fundamental frequency F0, [19], given as:
\[
\text{jitter} = \frac{1}{N} \frac{\sum_{i=1}^{N-1} |F_0_i - F_0_{i+1}|}{\sum_{i=1}^{N} F_0_i} \tag{5.19}
\]

Where \( N \) was the number of glottal cycles.

5.4.2.4 Shimmer (PS-S)

The shimmer measured the variation of the peak-to-peak speech wave amplitude from cycle-to-cycle. It was estimated as the first derivative of the peak-to-peak amplitude \( A \), [19], given as:

\[
\text{Shimmer} = \frac{1}{N} \frac{\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\sum_{i=1}^{N} A_i} \tag{5.20}
\]

Where \( N \) was the number of glottal cycles.

5.4.3 Prosodic features derived from the glottal waveform (PG)

Using the glottal waveform estimate (see Section 5.4.1), instead of the speech wave, the following prosodic features were calculated:

1. Fundamental frequency estimated from the glottal waveform (PG-F0)
2. Logarithmic energy of the glottal waveform (PG-LogE)
3. Jitter estimated from the glottal wave (PG-J)
4. Shimmer estimated from the glottal wave (PG-S)

The calculations were done using the same formulas as those described in Section 5.4.2.
5.4.4 Teager Energy Operator features (TEO)

Although, the classical source-filter model assumes a laminar air flow, recent laryngological experiments [50], [51], [75], suggest that the actual air flow during a glottal cycle has a nonlinear character which leads to the formation of air vortices between, and above the vocal folds. Acoustics effects of these formations can be manifested in generation of additional harmonic components and changes in spectral energy distribution. Zhou et al. [153], showed how these changes can be efficiently tracked using a parameter known as the area under the normalized Teager Energy Operator (TEO) autocorrelation envelope.

Variations of the TEO features were found to be particularly effective in speech based detection of stress [50], [51], [98], [153], and depression [75]. In all of these cases, the TEO parameters were calculated from the speech waveform.

In [153], a pairwise classification into neutral and angry speech, and loud and Lombard using the TEO features in provided accuracy ranging from 70% to 90%. Low et al. [75] compared the effectiveness of a wide range of acoustic speech parameters in the detection of clinical depression in adolescents, and showed that the TEO based parameters provide the best performance. The classification accuracy based on the TEO-CB-Auto-Env for female participants was up to 79% and for male participants up to 87%.

A number of earlier studies of speech based depression prediction [99], [100], pointed to the high importance of glottal features. This was consistent with similar observations described in [75], [84], [85], where the glottal features were reported to provide strong enhancement of the accuracy of depression diagnosis, especially when combined with prosodic or spectral features. To ameliorate the benefits of both, glottal
characteristics and the TEO sensitivity to nonlinear processes, the TEO method of feature extraction was applied in a traditional way to the speech waveform (producing the TEOS features), as well as to the glottal wave signal (producing the TEOG features).

The following sub-sections describe steps involved in the calculation of the TEOS and TEOG features.

5.4.4.1 Teager Energy Operator based features derived from speech waveform (TEOS)

For a discrete time domain signal samples $x[k]$, Kaiser [61], [62] proposed the following estimate of the speech instantaneous energy known as the Teager Energy Operator (TEO) [134], $\Psi$:

$$
\Psi(x[k]) = x^2[k] - x[k + 1]x[k - 1]
$$

(5.21)

The Teager energy operator (TEO) shows high sensitivity to transient changes in the signal instantaneous frequency and amplitude. As explained in [75], this property enables detection of additional harmonics generated by nonlinear airflow occurring periodically within each glottal cycle of the speech production process. Assuming that these nonlinearities are correlated with stress and emotions, Zhou et al. [153], investigated the use of TEO based features for stress and emotion classification in speech. As suggested in [153], if the speech signal is broken into auditory Critical Bands (CB) and the TEO parameters are calculated for each band, it is easier to observe the presence or absence of additional harmonic components within each band. Moreover, the speech analysis becomes more robust if the characteristic features are defined as the area under
the normalized TEO autocorrelation envelope (TEO-CB-Auto-Env). The normalized TEO autocorrelation function was defined as:

\[
R_{\psi(x)}[L] = \frac{1}{2M + 1} \sum_{n=-M}^{M} \psi(x[n])\psi(x[n + L])
\]  

(5.22)

where \(M\) was the number of samples within the analyzed speech frame and \(L\) is the correlation lag. Each speech frame was first separated into 15 critical bands using the Gabor band-pass filters with an impulse response function.

The TEO-CB-Auto-Env features were then calculated within the first 15 Critical Bands using procedures described in [75] and [153]. Figure 5.4 shows an example of the TEO energy profile followed by the autocorrelation envelope of a speech frame for a single critical band.

Table 5.3 shows the 15 critical bands generated from the Gabor band-pass filters and the frequency range for each critical band.

**Figure 5.4:** An example of the TEO-CB-Auto-Env feature extraction within the 6th critical band for a single speech frame: (a) TEO energy profile (b) Autocorrelation of the energy profile.
Table 5.3 Critical bands used to calculate the TEOS features

<table>
<thead>
<tr>
<th>Critical-band center frequency (Hz)</th>
<th>Bandwidth (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB_FC1 (200)</td>
<td>100-300</td>
</tr>
<tr>
<td>CB_FC2 (200)</td>
<td>300-500</td>
</tr>
<tr>
<td>CB_FC3 (200)</td>
<td>500-700</td>
</tr>
<tr>
<td>CB_FC4 (200)</td>
<td>700-900</td>
</tr>
<tr>
<td>CB_FC5 (200)</td>
<td>900-1100</td>
</tr>
<tr>
<td>CB_FC6 (200)</td>
<td>1100-1300</td>
</tr>
<tr>
<td>CB_FC7 (200)</td>
<td>1300-1500</td>
</tr>
<tr>
<td>CB_FC8 (400)</td>
<td>1500-1900</td>
</tr>
<tr>
<td>CB_FC9 (400)</td>
<td>1900-2300</td>
</tr>
<tr>
<td>CB_FC10 (400)</td>
<td>2300-2700</td>
</tr>
<tr>
<td>CB_FC11 (400)</td>
<td>2700-3100</td>
</tr>
<tr>
<td>CB_FC12 (400)</td>
<td>3100-3500</td>
</tr>
<tr>
<td>CB_FC13 (600)</td>
<td>3500-4100</td>
</tr>
<tr>
<td>CB_FC14 (600)</td>
<td>4100-4700</td>
</tr>
<tr>
<td>CB_FC15 (800)</td>
<td>4700-5500</td>
</tr>
</tbody>
</table>

5.4.4.2 Teager Energy Operator based features derived from glottal waveform (TEOG)

The glottal wave characteristics show high sensitivity to subtle voice changes and provide direct representation of the voice production process and onset of emotional expressions. These unique characteristics of the glottal domain were proven to be effective in providing important cues in detection of depression [75], [84], [85] and also in predicting the onset of depression [99], [100].

To combine the benefits of both, glottal characteristics and the TEO sensitivity to nonlinear processes, the TEO method of feature extraction was applied to the glottal wave extracted from the speech signal using the inverse filtering algorithm method.
Steps involved in the calculation of the TEO based glottal features (TEOG) are illustrated in Figure 5.6. The TEOG parameters were calculated within 15 auditory critical bands (CBs). The filtering process was applied to the glottal wave using a bank of band-pass Gabor filters with the centre frequencies and the band widths listed in Table 5.4. For each frame of the voiced speech, a set of features consisting of 15 TEOG coefficients representing areas under the normalised TEO autocorrelation envelopes of the glottal wave within 15 CBs was calculated as described in [100]. Table 5.4 also contains grouping of the CBs into 5 frequency bands B1–B5, with each band consisting of 3 CBs. As further explained in Section 5.8.3, the performance of the TEOG features calculated for the whole set of 15 bands was tested against TEOG performance within each of the individual CB groups B1-B5 in Section 5.8.4. Figure 5.5 shows an example of the TEO energy profile followed by the autocorrelation envelope of a glottal frame for a single critical band.

**Figure 5.5:** Example of the TEOG feature extraction within the 9th critical band for a single glottal frame: (a) TEOG energy profile (b) Autocorrelation of the energy profile
Figure 5.6: TEOG feature extraction implementation
Table 5.4 Critical bands’ (CB’) groups, CB’ center frequencies and bandwidths used to determine the TEOG features

<table>
<thead>
<tr>
<th>CB’ groups</th>
<th>Critical-band center frequency (Hz)</th>
<th>Bandwidth (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>CB_FC1 (200)</td>
<td>100-300</td>
</tr>
<tr>
<td></td>
<td>CB_FC2 (200)</td>
<td>300-500</td>
</tr>
<tr>
<td></td>
<td>CB_FC3 (200)</td>
<td>500-700</td>
</tr>
<tr>
<td>B2</td>
<td>CB_FC4 (200)</td>
<td>700-900</td>
</tr>
<tr>
<td></td>
<td>CB_FC5 (200)</td>
<td>900-1100</td>
</tr>
<tr>
<td></td>
<td>CB_FC6 (200)</td>
<td>1100-1300</td>
</tr>
<tr>
<td>B3</td>
<td>CB_FC7 (200)</td>
<td>1300-1500</td>
</tr>
<tr>
<td></td>
<td>CB_FC8 (400)</td>
<td>1500-1900</td>
</tr>
<tr>
<td></td>
<td>CB_FC9 (400)</td>
<td>1900-2300</td>
</tr>
<tr>
<td>B4</td>
<td>CB_FC10 (400)</td>
<td>2300-2700</td>
</tr>
<tr>
<td></td>
<td>CB_FC11 (400)</td>
<td>2700-3100</td>
</tr>
<tr>
<td></td>
<td>CB_FC12 (400)</td>
<td>3100-3500</td>
</tr>
<tr>
<td>B5</td>
<td>CB_FC13 (600)</td>
<td>3500-4100</td>
</tr>
<tr>
<td></td>
<td>CB_FC14 (600)</td>
<td>4100-4700</td>
</tr>
<tr>
<td></td>
<td>CB_FC15 (800)</td>
<td>4700-5500</td>
</tr>
</tbody>
</table>

5.4.5 Spectral features derived from speech waveform (SS)

Spectral features included various parameters characterizing properties of the speech power spectrum (SPS). Spectral properties have been previously tested in speech based classification of diagnosed depression, providing moderately accurate results [39], [75], [85]. The spectral features investigated in this study included:

1) Spectral flux,
2) Spectral centroid,
3) Spectral entropy,
4) Spectral roll-off rate,
5) Power spectral density (PSD) and
6) Formant values.
These parameters were derived from the fast Fourier transform (FFT) applied to the speech wave on the frame by frame basis. Spectral features in previously reported depression recognition tests [75], provided about 69% of accuracy for both male and female depressed adolescents.

The following sub-sections describe the spectral feature parameters with more details.

5.4.5.1 Spectral flux

Spectral flux measures the degree of change between the power spectrum across consecutive frames. It was calculated as the Euclidean distance between the normalized power spectrum of a given frame, $k$ and the normalized power spectrum of the previous frame, $k-1$.

$$ Flux = |PS(k) - PS(k - 1)| $$  \hspace{1cm} (5.23)

5.4.5.2 Spectral centroid

Spectral centroid measures the center point of the spectral power within the frequency distribution. It was calculated as the weighted mean of frequencies distribution of the signal spectrum with their spectral magnitudes acting as weights.

$$ Centroid = \frac{\sum_{0}^{N-1} PS(n)f(n)}{\sum_{0}^{N-1} PS(n)} $$  \hspace{1cm} (5.24)

N represents the total number of frequencies, $f(n)$ with the corresponding power spectrum magnitudes, PS(n).
5.4.5.3 Spectral entropy

Spectral entropy measures the average amount of Shannon’s spectral information calculated for the normalised spectral amplitudes within each frame. Let \( NPS(n) \) be the normalized spectrum, thus the entropy was given as:

\[
Entropy = - \sum_{n=1}^{M} NPS(n) \log_2 NPS(n)
\]  

(5.25)

5.4.5.4 Spectral roll-off

Roll-off characterises signal's distribution of energy vs frequency. For a given frame, the spectral roll-off was determined as the frequency bin number below which 75\% of the total spectral energy was accumulated.

\[
Rolloff = 0.75 \sum_{n=0}^{N-1} |PS(n)|
\]  

(5.26)

5.4.5.5 Power spectral density (PSD)

The PSD were computed using the Welch spectral estimator method. The PSD features for each frame included: the total power for the whole bandwidth (0-2000 Hz), powers of four sub-bands (0-500Hz, 500-1000Hz, 1000-1500Hz, and 1500-2000Hz) and the ratios of powers for each spectral sub-band to the total power of the whole bandwidth.

5.4.5.6 Formants

Formants are the resonant frequencies of the vocal tract. They are known to carry essential information necessary for understanding speech [124]. Each peak of the vocal
tract spectrum indicates the formants, corresponding to frequency and -3dB bandwidth. For each frame, the first three formant frequencies (FMT1, FMT2 & FMT3) and their bandwidths (FBW1, FBW2 & FBW3) were estimated from the spectral envelope given by the 13th order LP filter.

5.5 Gaussian Mixture Model (GMM) and Bayesian Classification

In all single-channel components of speech classification systems described in this chapter, features extracted from the speech data were used to generate models of two acoustic classes AR and NAR using the Gaussian Mixture Model (GMM) method. The GMM modeling (or training) stage was in the testing stage integrated with the Bayesian classification decision procedure which determined the most probable class for the test (or query) samples.

The GMM is a widely recognized classifier scheme and has been effectively used in modeling speech information for recognition tasks [110]. Gaussian distributions are commonly used to model the distributions of continuous random variables. In order to model a dataset category for the purpose of this study, a linear combination GMM of order M models was used, where the probability density function of a data vector $x$ was calculated as a weighted sum (or mixture) of M different Gaussian densities shown in Equation (5.27). Each Gaussian density $N(x|\mu_m, \Sigma_m)$ has its own weight $\pi_m$, mean $\mu_m$ and covariance $\Sigma_m$. Each mixture for the Gaussian densities are normalized to be $0 \leq \pi_m \leq 1$. Diagonal covariance matrices were used instead of full covariance for finding the Gaussian components to avoid over fitting where small amounts of data are available.
The expectation maximization (EM) algorithm is commonly used to estimate the optimal values of these parameters by maximizing the expected log likelihood function given in Equation (5.28), where \( X = \{ x_1, \ldots, x_N \} \).

\[
\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left( \sum_{m=1}^{M} N(X_n|\mu_m, \Sigma_m) \right)
\]  

(5.28)

This is known as the E-step which is the first stage of the EM algorithm in finding the expected parameters using latent values. The next stage of the EM algorithm is the M-step which operates to maximize the parameters of the log likelihood function following Equation (5.29) and re-estimating the E-step. The EM iteratively improves the estimation of the log likelihood function by repeating the process of maximizing the log likelihood function found and re-estimating the parameters (i.e. alternating between the E-step and M-step) until convergence is met.

\[
\argmax_{\pi, \mu, \Sigma} \ln p(X|\pi, \mu, \Sigma)
\]  

(5.29)

The Gaussian mixture modeling or training stage is usually integrated with the Bayesian classification decision procedure [16], [40], [53], [135] which determines the most probable class for a given query sample. A 3rd order Gaussian mixture model combined with the forward-backward EM algorithm was trained to generate models for
the two acoustic classes AR and NAR for our single-channel classification experiments. The HTK toolbox [151] was used to implement the Gaussian mixture modeling process illustrated in Figure 5.7 and the Bayesian classification shown in Figure 5.8. Both the modeling (training) and the classification (testing) stages required prior knowledge of classes to which the training or testing speech samples belonged.

Following the Bayes formula shown in Equation (5.30), the classification of a testing sample can be done by evaluating the probability, \( P(c_k|x) \) of a testing feature vector \( x \) belonging to a class \( c_k \) of \( k \) different classes (AR and NAR)

\[
P(c_k|x) = \frac{p(x|c_k)p(c_k)}{p(x)}
\]  

(5.30)

where \( p(x|c_k) \) is the probability density function of class \( c_k \) and \( P(c_k) \) is the priori probability of the class.

### 5.6 Normal Distribution Tests

The signal classification framework undertaken in this study was based on the underlying assumption that the analyzed data had a normal (Gaussian) distribution. In order to justify the validity of our experiments, the characteristic features were examined to verify normal distribution within each of the two classes “at risk” (AR) and “not at risk” (NAR). The Kolmogorov-Smirnov (KS) test [38] provided positive results indicating that all feature coefficients listed in Table 5.2 were normally distributed where the alternative hypothesis was rejected.

After confirming normal distributions of data, two further statistical tests MANOVA and ANOVA were conducted in order to determine which of the feature
parameters show a significant level of correlation with the two states: AR and NAR of depression. The aim was to eliminate from further analysis features that did not show a significant correlation.

The multivariate analysis of variance (MANOVA) with the Wilk’s lambda statistical procedure was used to compare multivariate means of the feature categories listed in Table 5.2. The means were compared between two classes AR and NAR, and for each feature category a separate MANOVA test was performed as recommended in [38]. Separate testing of each category had the additional benefit of validating the overall integrity of feature combinations included in each category. The results indicated that all the main feature categories showed a significant difference (p<0.05) between means for the two classes AR and NAR.

The multivariate MANOVA analysis was followed by the univariate, one-way analysis of variance (ANOVA) to determine the individual significance levels for each of the feature sub-categories listed in Table 5.2. The same criterion of p<0.05 was used to indicate statistically significant results. Results obtained from the ANOVA test indicated that all feature sub-categories except jitter showed significant (p<0.05) differences of mean values between the AR and NAR state. The ANOVA result for jitter was p=0.176 indicating that the mean values of jitter for the AR and NAR groups were not significantly different, therefore in further analysis the jitter features were eliminated from the prosodic category including both PS and PG features.
5.7 Experimental Setup for Speech Modeling and Classification

To validate the capability of a reliable classification system for the prediction of major depression from speech signals of non-depressed adolescents, four different types of classification systems were tested and compared to determine the effectiveness and reliability of the prediction methodology.

The classification methodologies tested in this chapter and listed in the order of an ascending developmental complexity were:

1) Classical, single-channel speech classification (SCSC) system;
2) New, multi-channel speech classification (MCSC) system
3) New, multi-channel weighted speech classification (MCWSC) system;
4) New optimised multi-channel weighted speech classification (OMCWSC) system.

Sections 5.7-5.11 explain in detail the testing/validation setups and procedures for these systems.

All systems were tested using the same dataset of speech recordings collected from 30 participants (15 AR and 15 NAR) as described in Section 5.2. Although the database contained speech recorded during Problem Solving Interaction (PSI) and Event Planning Interaction (EPI) sessions, the experiments used only speech recorded during the PSI sessions. It was motivated by the fact that the PSI recordings were previously shown to be more effective in depression classification than the EPI recordings [75], [101]. One of the possible reasons is that the PSI sessions were more likely to evoke conflictual behavior leading to verbal displays of negative emotions which are known to have strong correlation with adolescents’ depression in family environments [128].
For a single-channel system tested on its own, approximately 50% of the dataset samples were used for training and the remaining 50% for testing. In the case of a multi-channel system, 50% of the dataset samples were used for training and the remaining 50% were divided into two equal size sub-sets of speech data samples; one of these sub-sets was used to determine the weights for the multi-channel classification system and the other one was used to test the multi-channel system. All results were averaged over 3 stratified cross validation runs, with each run performed on different and mutually exclusive sub-sets of training, weight calculation and testing data. In the case of NAR samples, for each cross validation case, 15 participants were randomly chosen from the total pool of 176 NAR participants and then 8 (or 7) participants were randomly chosen out of 15 for the training and the remaining 7 (or 8) were used in the testing process. In the case of AR participants, for each cross-validation run, 8 (or 7) participants were randomly chosen from the total pool of 15 participants for the purpose of training and the remaining 7 (or 8) were used in the testing process.

The predictive classification results were assessed based on three statistical parameters: sensitivity, specificity and accuracy given by Equations (5.31), (5.32) and (5.33) respectively. These parameters are commonly used in the assessment of diagnostic methods and binary classification algorithms [75], [85], [10].

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\% \quad (5.31)
\]
Specificity = \frac{TN}{TN + FP} \times 100\% \quad (5.32)

Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (5.33)

The TP, FP, FN and TN parameters in Equations (5.31)-(5.33) were calculated using two separate approaches: utterance based (UB) and person based (PB). For these two approaches the following definitions of TP, FP, FN and TN applied:

- TP was the number of true positive outcomes defined either as the number of AR adolescents (PB approach) or utterances (UB approach) classified as AR.
- FP was the number of false positive outcomes defined either as the number of NAR adolescents or utterances classified as AR.
- TN was the number of true negative outcomes defined either as the number of NAR adolescents or utterances classified as NAR.
- FN was the number of false negative results defined either as the number of AR adolescents or utterances classified as NAR.

For the utterance based (UB) approach, the number of utterances classified correctly or incorrectly as AR or NAR were calculated as in Equation (5.34). In the case of the person based (PB) approach, the numbers of adolescents classified correctly or incorrectly as AR or NAR were calculated as shown in Equation (5.35). An adolescent was then classified as AR if more than 50% of utterances representing this adolescent were correctly classified as AR. Similarly, an adolescent was classified as NAR if more than 50% of utterances representing this adolescent were correctly classified as NAR.
When assessing the performance, a well performing system would have high values for all of these three parameters (sensitivity, specificity and accuracy), however if a compromise had to be made, it was desired for the sensitivity (AR adolescents correctly diagnosed as AR) to be slightly higher than the specificity (NAR adolescents correctly diagnosed as NAR). This way a safer depression screening assessment could be made without missing too many true-positive cases.

The 50% criterion used in the PB approach was experimentally tested to provide high TP rates while still preserving reasonable FP rates. As the predictive diagnostic method for depression was expected to have relatively large true positive (TP) rates (i.e. large sensitivity), at the cost of lower values of false positive (FP) rates, the chosen threshold ratio of 50% provided best fit into these criteria and was also previously found to give the best PB classification results [101] in depression prediction experiments.

The processed speech data contained a relatively large amount of data representing each person; however the total number of persons was relatively small. The PB approach was therefore undertaken to mitigate the underestimation of the person-independent (UB) classification due to a small number of participants. The study of the
person dependent (PB) method was also important for future depression prevention programs based customised depression prediction approaches, where the data collected from individual patients could be used to design customised risk assessment algorithms.

5.8 Classical, Single-Channel Speech Classification (SCSC) System

A single-channel speech classification (i.e. using only one type of features and a single modelling method) was conducted to evaluate the performances of individual feature categories, shown in Table 5.2, in the prediction of clinical depression. Each individual feature category (G, PS, PG, TEOS, TEOG and SS) provided an independent input to the single-channel classification system.

Experiments with the SCSC approach first tested the acoustic features derived from the speech waveforms, this was followed by features extracted from the glottal waveforms and finally, the TEOG features calculated within different frequency sub-bands (TEOG-SB1 - TEOG-SB5) were tested.

In order to model the individual training set of each feature category for the “at risk” (AR) and “not at risk” (NAR) classes, the single-channel modelling and classification adopted the Gaussian Mixture Model (GMM) method. A Bayesian classification was used to determine the prediction outcome for the final decision making of each testing input based on the modeled classes from the GMM model.

The single-channel classification experiments were conducted in order to:

1) Determine if the risk for Major Depression can be predicted from speech signals of adolescents 2.5 years before the full blown depression symptoms (detectable by classical diagnostic methods) become apparent;
2) Determine the effectiveness of different types of acoustic features of speech in
discrimination between adolescents that are “At Risk” (AR) and “Not At Risk”
(NAR) of developing Major Depression within the next 2.5 years.

In addition, test conducted with the single-channel approach provided fundamentals for
the design and implementation of new, more advance multi-channel classification
systems (MWSC and OMWSC), where the outputs of individual single-channel
classifiers can be evaluated and used to derive a prediction result based on weighted
combination of outcomes provided by different types of features.

The following sub-sections describe the single-channel experimental framework
and the classification (prediction) results.

5.8.1 General Framework of the Single Channel Speech Classification (SCSC) System

Figure 5.7 & Figure 5.8 illustrate the overall framework of the basic single-channel
prediction procedure commonly used in speech classification and recognition systems
[39], [49], [50], [75], [84], [85]. The procedure included two main stages: training stage
(see Figure 5.7) and classification and decision making stage (see Figure 5.8). The
training stage used a set of data with known classes to build statistical models of the AR
and NAR classes. After the pre-processing, feature parameters characterizing the speech
acoustics of each class were calculated. The probability density functions of these
parameters were passed to the Gaussian Mixture Model (GMM) algorithm as explained
in the previous section, to build the Gaussian Mixture Models [151] of the AR and NAR
classes. For all different types of acoustic speech parameters listed in Table 5.2, separate
GMMs were built to represent each class (AR and NAR) for each different acoustic
feature categories. These class models were then used in the classification stage as shown in Figure 5.8.

Procedures involved in the classification and decision making process for the testing stage also included pre-processing, feature extraction and Gaussian density calculations. This was followed by a relatively simple process of Bayesian classification [16], [40], [53], [135], that evaluates the likelihood probability of each testing sample to the AR and NAR classes generated from the GMM models in order to provide the final class estimate (prediction result).

---

**Figure 5.7:** An overview of the training stage of single-channel system generating AR and NAR models for a given type of features (G, P, TEO or S).

**Figure 5.8:** An overview of a single-channel classification system using models from the training stage.
5.8.2 Testing the SCSC system with G, PS, TEOS and SS features

The prediction performance given by four single-channel classification techniques with each technique using a single type of features was tested to evaluate and compare the effectiveness of an early prediction of clinical depression in adolescents using glottal features (G) and features derived from the speech waveform (PS, TEOS and SS). It also provided a background for selection of features most suitable for the new multi-channel techniques discussed in the later sections of this chapter.

The performance of each feature category in predicting depression was evaluated in terms of three parameters: sensitivity, specificity and accuracy explained in Section 5.7. Table 5.5 and Table 5.6 show results provided by each type of features when using the utterance based (UB) approach and when using the person based (PB) approach, respectively. The UB and PB approaches were described in Section 5.7.

It is likely that the lesser performance provided by the utterance based (UB) approach when compared to the person based (PB) approach was due to the relatively small number of participants representing each class (15 for AR and 15 for NAR). These numbers were too small to build person-independent class models. However; as Table 5.6 shows, the person based (PB) approach lead to quite satisfactory results. This was due to the fact that each person was represented by a relatively large number of speech samples which means that the number was sufficient to build adequate statistical representations of each individual.

As seen in Table 5.5, the highest yield of overall classification accuracy using the UB approach was just approximately 5% above the chance prediction level for the prosodic features followed by 2.6% for the glottal features and 1.4% for the spectral
features. The TEOS feature category was 0.7% below the 50% chance prediction level. In terms of sensitivity to specificity ratio, all features except the TEOS was able to produce a sensitivity which is slightly higher than the specificity.

**Table 5.5: Depression Prediction Results for Individual Features from Speech Waveform Using the Utterance Based (UB) Approach**

<table>
<thead>
<tr>
<th>Features</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glottal</td>
<td>56.43</td>
<td>48.73</td>
<td>52.58</td>
</tr>
<tr>
<td>Prosodic</td>
<td>62.19</td>
<td>47.75</td>
<td>54.97</td>
</tr>
<tr>
<td>TEO</td>
<td>45.17</td>
<td>53.36</td>
<td>49.27</td>
</tr>
<tr>
<td>Spectral</td>
<td>55.83</td>
<td>46.93</td>
<td>51.38</td>
</tr>
</tbody>
</table>

**Table 5.6: Depression Prediction Results for Individual Features from Speech Waveform Using the Person Based (PB) Approach**

<table>
<thead>
<tr>
<th>Features</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glottal</td>
<td>76.19</td>
<td>62.5</td>
<td>69.35</td>
</tr>
<tr>
<td>Prosodic</td>
<td>73.21</td>
<td>52.98</td>
<td>63.1</td>
</tr>
<tr>
<td>TEO</td>
<td>43.45</td>
<td>60.71</td>
<td>52.08</td>
</tr>
<tr>
<td>Spectral</td>
<td>60.12</td>
<td>47.62</td>
<td>53.87</td>
</tr>
</tbody>
</table>

Table 5.6 shows that the glottal features provided the best performance with accuracy close to 70% and a desirable sensitivity to specificity ratio of 76%/63% which means that this ratio met our preference criterion of sensitivity being slightly higher than the specificity. The prosodic features were slightly worse in terms of accuracy (63%) and the specificity to sensitivity ratio of 73%/53% was also less desirable as the specificity was low. Both spectral and TEOS features provided poor prediction results with all three parameters (specificity, sensitivity and accuracy) oscillating around the baseline of 50% representing a chance level.
In summary, the depression prediction tests performed on individual types of features derived from speech waveform showed that only the glottal (G) and prosodic (PS) features could be considered as reasonably effective on their own in predicting depression with a desirable specificity/sensitivity ratio and accuracy significantly higher than the chance level. This also highlighted the importance and capability of the glottal features in the prediction of clinical depression which provided satisfactory results. In view of this, our experiments based on the single-channel approach extended on investigating features from the glottal domain which carries some distinguishing information between adolescents who are “At Risk” of depression from those who are “Not At Risk” of depression.

Importantly, with the single-channel prediction accuracy of 69% for the glottal (G) features, the results firmly supported the idea that acoustic speech analysis and classification can be used to predict risk for depression in adolescents.

As described in the following sections, the highest performance achieved for the glottal features prompted further investigations of these features in single-channel and then in multi-channel systems.

It was also shown that the person based (PB) prediction offered more suitable approach when compared to the utterance based (UB) method.

5.8.3 Testing the SCSC System with TEOG and PG features

As shown in the previous section, the glottal features (G) clearly outperformed features derived from the speech waveform (PS, TEOS and SS) specifically when using the PB approach. For this reason, further experiments investigating the performance of the SCSC
system with features based on the glottal wave and the PB approach were conducted. Two types of features were tested: the Teager Energy Operator features derived from glottal waveform (TEOG) and the prosodic features also derived from glottal waveform (PG).

Table 5.7 shows the classification results for the TEOG and PG features within the PB classification framework.

It can be observed that the TEOG features provided better classification performance outperforming the PG features on average by 3.6% across all three different evaluation measurements (sensitivity, specificity and accuracy). When comparing these results with the results in Table 5.6, the achieved accuracy of 65% for TEOG and 61% for PG was below the 69% accuracy previously achieved for the glottal features (G).

There was a clear improvement for the TEOG features derived from the glottal waveform when compared to TEO features derived from the speech waveform (TEOS). An average increase of 25% in terms of sensitivity and 12.5% in accuracy was achieved for TEOG with no change in specificity which was 60.71% in both cases.

When comparing the prosodic feature derived from glottal wave (PG) in Table 5.7 to the prosodic features derived from the speech wave (PS) in Table 5.6, there was a decrease in accuracy by 2.1% and 8.33% in sensitivity for the PG features whereas the specificity for the PG was slightly increased by 4.16%.

In summary, it was observed that the tested features derived from glottal waveform (TEOG and PG) did not match the prediction performance of the glottal features (G). However, high performance of the G features on their own, as well as the improvement of the TEO features derived from the glottal wave (TEOG) when compared
to the performance of the TEO features derived from the speech wave provided an experimental evidence strongly supporting the importance of glottal parameters in for the detection and prediction of depression. These observations are consistent with similar findings reported in [75], [84], [85].

With the TEOG feature providing promising prediction performance, further investigations into the frequency dependent nature of the TEOG were conducted. These experiments are discussed in the next section.

Table 5.7: Depression Prediction Results for Individual Features from Glottal Waveform Using the Person Based (PB) Approach

<table>
<thead>
<tr>
<th>Features</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEOG</td>
<td>68.45</td>
<td>60.71</td>
<td>64.58</td>
</tr>
<tr>
<td>PG</td>
<td>64.88</td>
<td>57.14</td>
<td>61.01</td>
</tr>
</tbody>
</table>

5.8.4 Testing the SCSC system with the TEOG features calculated within different frequency sub-bands (TEOG-SB1 – TEOG-SB5)

As suggested in [48], [50], [112], contributions of the TEO features (TEO-CB-Auto-Env) in stress detection and to the prediction of depression [100], vary across different frequency sub-bands. Frequency dependent nature of the AM-FM features were also investigated for a speech based emotion classification in [126]. In the previous section, the TEOG features representing all frequency bands (100Hz to 5500Hz) were investigated. This section describes experiments aiming to determine how the different frequency sub-bands of the TEOG features contribute to the prediction of depression in adolescents.
As explained in Section 5.4.4, the 15 TEOG coefficients corresponded to the TEOG estimates within each of the 15 Critical Bands (Table 5.3) covering the frequency range from 100Hz to 5500Hz.

In this section the TEOG coefficients were tested in 5 separate groups (TEOG-SB1-TEOG-SB5) of 3 coefficients and one group of 9 coefficients (TEOG-SB3-5). These groups were generated to represent the TEOG coefficients within the following frequency sub-bands:

1) TEOG-SB1: 100Hz-700Hz
2) TEOG-SB2: 700Hz-1300Hz
3) TEOG-SB3: 1300Hz-2300Hz
4) TEOG-SB4: 2300Hz-3500Hz
5) TEOG-SB5: 3500Hz-5500Hz
6) TEOG-SB(3-5): 1300Hz-5500Hz

These groupings were made in order to assess and compare the performances of TEOG features within low, medium and high frequency ranges.

Table 5.8 presents classification results provided by each individual frequency sub-band of the TEOG features.

It can be observed that generally the overall classification accuracy is higher for the high frequency sub-bands SB3, SB4 and SB5 (above 60%) than for the low frequency sub-bands SB1 and SB2 (below 60%).

The highest prediction accuracy of almost 68% was achieved for the SB4 alone (2300Hz – 3500Hz) and the lowest accuracy of 52% was achieved for the SB2 (700Hz – 1300Hz).
The accuracy of the combined high frequency bands SB3, SB4 & SB5 (1300Hz - 5500Hz) provided the same accuracy of 68% as the SB4 alone. Interestingly, this result was higher than the classification accuracy of about 65% achieved for the whole bandwidth (100Hz – 5500Hz), this is when using all sub-bands together (Table 5.7).

These results clearly indicate that the TEOG features calculated within the frequency band ranging from 2300Hz to 3500Hz provide the strongest discrimination between speech characteristics of AR and NAR adolescents.

In summary, the results showed that the TEOG features provide better discrimination between the AR and NAR classes when evaluated at higher frequency levels (i.e. frequencies higher than 1300Hz). In particular the TEOG features estimated within the frequency range from 2300Hz to 3500Hz were found to be the most effective indicators of the risk for depression. As suggested in [153], if the speech signal is broken into auditory Critical Bands (CB) and the TEO parameters are calculated for each band, it is easier to observe the presence or absence of additional harmonic components within each band. In [75], the correlation between TEO and Glottal features relating to additional harmonics found in speech of depressed subjects explains the advantage of these features as indicators of depression. Here, the TEOG feature showed that higher frequency bands provided better classification accuracy could be explained by the fact that the presence of additional harmonics are more evident in these frequency levels. This however should be further investigated to determine the strength of the produced additional harmonics in both classes and also in different frequency bands.
Table 5.8: Depression Prediction Results for Individual TEOG Sub-Band Features Using the Person Based (PB) Approach

<table>
<thead>
<tr>
<th>TEOG sub-band features</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB1 (CB1-CB3)</td>
<td>55.95</td>
<td>52.98</td>
<td>54.46</td>
</tr>
<tr>
<td>SB2 (CB4-CB6)</td>
<td>60.12</td>
<td>44.05</td>
<td>52.09</td>
</tr>
<tr>
<td>SB3 (CB7-CB9)</td>
<td>64.88</td>
<td>61.90</td>
<td>63.39</td>
</tr>
<tr>
<td>SB4 (CB10-CB12)</td>
<td>64.29</td>
<td>70.83</td>
<td>67.56</td>
</tr>
<tr>
<td>SB5 (CB13-CB15)</td>
<td>68.45</td>
<td>66.07</td>
<td>67.26</td>
</tr>
<tr>
<td>TEOG (SB3-SB5)</td>
<td>66.07</td>
<td>69.05</td>
<td>67.56</td>
</tr>
</tbody>
</table>

5.9 New, Two-Stage Multi-Channel Speech Classification (MCSC) System

Building up from Section 5.8, where the classical single channel classification of individual speech features categories was tested, in this section the investigation moves into a new multi-channel speech classification approach. Various methods of multi-modal systems relating to emotion recognition were reviewed in [72]. This includes fusion at feature level, decision level, meta level and hybrid fusion. It was concluded that there is no definite answer to which fusion method best improves a single channel approach. The multichannel approaches proposed in this thesis focuses on decision level fusion. The intention of investigating on decision level was to evaluate the performance of individual single channel feature categories and to find an optimal combination for these single channels. A two-stage SVM was investigated in [59], for depression diagnosis and monitoring, however the proposed approach here uses a two-stage GMM classification for the prediction of depression.

The proposed multi-channel speech classification (MCSC) system, as illustrated in Figure 5.9, is a two-stage procedure. These two stages can be described as follows:
Stage 1: Individual feature categories are first classified using $N$ parallel single-channel (SCSC) classification approaches based on the GMM classifier as described in Section 5.8. Class estimates $\hat{y}_1, \ldots, \hat{y}_N$ generated by each of the $N$ channels are combined to form a feature matrix $M = [\hat{y}_1, \ldots, \hat{y}_N]$ to be used to train and test the second stage of the classification procedure.

Stage 2: Feature matrices $M$ obtained during the 1st stage are used to conduct another single-channel (SCSC) classification based on the GMM classifier. The classification outcome from the second stage provides the final prediction decision (AR or NAR).
Figure 5.9: Overview of the two-stage GMM multi-channel classification system

5.9.1 Testing the MCSC system with the G, PS, TEOS & SS features

The MCSC, two-stage GMM multi-channel speech classification approach, was trained and tested using four individual feature categories (four channels): glottal (G), prosodic derived from the speech wave (PS), TEO derived from the speech wave (TEOS) and spectral characteristics of the speech wave (SS).
During the testing procedure, input query samples $x_t$ of speech were classified by the four parallel channels in the 1st stage of the MCSC process. The classification results were then used to form feature matrices $M(x_t)$ which were used in the second stage to provide the final prediction result.

Table 5.9 shows the test results for the two-stage MCSC system. The sensitivity/specificity ratio of 58%/56% and accuracy of 57% were in this case not satisfactory as they did not reach the average levels obtained for the single-channel approaches (see Table 5.5-Table 5.8). The overall accuracy for the two-stage MCSC system was only by 5% higher than the TEOS and by 3% higher than SS performance. In comparison with other cases, the MCSC system showed lower performance.

In conclusion, the two-stage MCSC approach introduced an interesting concept of a multi-channel classification with each channel performing independent assessment based on different types of features, however the idea of combining the outcomes of individual channels into a secondary set of classification features did not deliver good results.

The following sections are investigating more efficient ways of the final decision making process for the multi-channel speech classification system.

<table>
<thead>
<tr>
<th>MCSC feature combination</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G, PS, TEOS, SS</td>
<td>58.33</td>
<td>55.56</td>
<td>56.95</td>
</tr>
</tbody>
</table>
5.10 New, Multi-Channel Weighted Speech Classification (MCWSC) System

In this section, a new multi-channel weighted speech classification (MCWSC) approach is described.

The overall concept of the MCWSC is shown in Figure 5.10 and Figure 5.11. At the first stage, speech samples were classified as AR or NAR using a parallel configuration of 4 single-channel classifiers (described in Section 5.8). Each channel included a single Bayesian classifier and two Gaussian mixture models (one for AR and one for NAR) developed using a single category of features (e.g. G, PS, TEOS or SS described in Section 5.4).

At the second stage, a subset of the outputs from the 4 single-channels were used to perform a supervised weights allocation and the remaining subset of the 4 single-channel outputs were combined into a final decision by calculating a weighted sum of the intermediate decisions generated by each channel.

Two approaches to the classification process were investigated:

a) Classification with a single set of weights (1W), and

b) Classification with two sets of weights (2W)

5.10.1 Finding the weight values

As illustrated in Figure 5.10, class estimates produced by each single-channel classification system were used to calculate the weight values for the final classification decision procedure shown in Figure 5.11.

Values of the weight coefficients were determined using a supervised classification process. A sub-set of the speech data was used for this purpose, and
differed to the sub-sets used for the GMM training of the models and the multi-channel classification process. The supervised character of this process meant that for each data sample $x_i$ ($i=1,\ldots, L$), the actual class $y(x_i)$ to which this sample belonged, was known, and therefore the classification accuracy corresponding to the class estimates produced by the system could be assessed. This multi-channel training data was classified separately by each of the 4 channels producing four sets of class estimates $\tilde{y}_k(x_i), k=1,\ldots,4$.

These class estimates given by each channel were used to calculate the values of the objective function given as:

$$f_{obj}(k) = \frac{1}{L} \sum_{i=1}^{L} (\tilde{y}_k(x_i) - y_k(x_i))^2$$

(5.36)

where $k=1,\ldots,N$ is the channel number, $L$ is the total number of the speech samples, $y_k(x_i)$ is the actual class of speech sample $x_i$ and $\tilde{y}_k(x_i)$ is the class estimate given by the $k$th channel. The actual class labels and the estimated class labels were set to be equal to +1 for the AR class and -1 for the NAR class.

The objective function $f_{obj}(k)$ defined in Equation (5.36) represents a total average squared error between the actual classes and the classes estimated by the classification channel $k$. It is therefore a measure of the classification accuracy given by the feature category used by this channel.

To be able to compare the channels and the associated features performance, normalized values of the objective function were calculated as follows:

$$W_k = \frac{1}{\sum_{k=1}^{N} \left( \frac{1}{f_{obj}(k)} \right)}$$

(5.37)
The normalization process provided \( k \) weight parameters \( W_k \), which met the following property:

\[
\sum_{k=1}^{N} W_k = 1
\]  

(5.38)

The described process of weight calculation assigned the higher weight values to the classification channels (i.e. the corresponding feature category), that provided better classification results and smaller values to those that provided lesser performance. The additional advantage given by this approach was that by comparing the relative performance of different types of features parameters, the effects of clinical depression on speech acoustics and to a certain degree the physiology of phonation can be observed.

For the purpose of comparison, the values of weight parameters were determined using two different approaches. In the first instance, a single set of weights (1W) containing equal amounts of samples representing both classes AR and NAR was derived. In the second instance, two sets of weights (2W) were derived, one using only speech data representing the AR group (\( W_{AR} \)) and one using speech data representing only the NAR group (\( W_{NAR} \)).
Figure 5.10: Determining the weight values for the MCWSC classification decision

5.10.2 Generating the decision matrix for the MCWSC system

After finding the values of the weight coefficients, the system classification (prediction) performance was tested as follows.

Input speech samples denoted in Figure 5.11 as query samples were first classified by each of the 4 channels and each channel produced its own class estimates \( \tilde{y}_k(x_i) \).

The class estimates given by the channels were combined into a weighted score parameter \( r(x_i) \) given as:
\[ r(x_i) = \sum_{k=1}^{N} W_k \tilde{y}_k(x_i) \]  

(5.39)

where, \( k = 1, \ldots, N \) is the channel number, \( W_k \) is the weight parameters and \( \tilde{y}_k(x_i) \) is the class estimate given by the kth channel.

The final classification decision was then made based on the sign of \( r(x_i) \).

Two different final classification decision matrices illustrated in Table 5.10 and Table 5.11 were used depending on the number of calculated weight coefficients.

**Table 5.10: Decision Matrix for a Single Set of Weights (1W)**

<table>
<thead>
<tr>
<th>sign of ( r(x_i) )</th>
<th>Final prediction result</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>AR</td>
</tr>
<tr>
<td>-</td>
<td>NAR</td>
</tr>
</tbody>
</table>

**Table 5.11: Decision Matrix for a Two Set of Weights (2W)**

<table>
<thead>
<tr>
<th>sign of ( r_{AR}(x_i) )</th>
<th>sign of ( r_{NAR}(x_i) )</th>
<th>If condition</th>
<th>Final prediction result</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>Undetermined</td>
<td>AR</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>(</td>
<td>r_{AR}(x_i)</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>(</td>
<td>r_{AR}(x_i)</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>(</td>
<td>r_{AR}(x_i)</td>
</tr>
</tbody>
</table>

If a single set of weights (1W) was calculated, then a simple decision matrix for the \( r(x_i) \) values shown in Table 5.10 was used. When two sets of weights were calculated, two different values of \( r(x_i) \) were produced for the same speech sample, \( x_i : r_{AR}(x_i) \) and \( r_{NAR}(x_i) \) for the weights derived from the multi-channel training data representing the AR.
and NAR speech samples, respectively. The final class was then assigned using a set of rules presented in Table 5.11 (a positive or negative sign indicated AR or NAR, respectively).

**Figure 5.11**: The multi-channel classification system using MCWSC system weights
5.10.3 Testing the MCWSC system with G, PS, TEOS and SS features

The MCWCS tested four types features of speech features (G, PS, TEOS and SS). This same set of features was used for consistency to be able to compare the prediction capability of the MCWSC with the test results for the SCSC (Section 5.8.2) and MCSC (Section 5.9.1).

Although the TEOS and SS features were rather weak predictors of depression on their own (see Table 5.5 and Table 5.6), it was possible that they could provide better performance within a multi-channel system that assigns different weights to the decisions based on individual feature performances.

Approximately 50% of the dataset samples were used for training. The remaining 50% were divided into two equal size sub-sets of speech data samples. One of the sub-sets was used to determine the weights for the multi-channel classification system and the other one was used to test the multi-channel system. All results were averaged over 3 stratified cross validation runs, with each run performed on different and mutually exclusive sub-sets of training, weight calculation and testing data.

In the case of NAR samples, for each cross validation case, 15 participants were randomly chosen from the total pool of 176 NAR participants and then 8 (or 7) participants were randomly chosen out of 15 for the training and the remaining 7 (or 8) were used in the testing process. In the case of AR participants, for each cross-validation run, 8 (or 7) participants were randomly chosen from the total pool of 15 participants for the purpose of training and the remaining 7 (or 8) were used in the testing process.

As described in Section 5.10.1 and 5.10.2, at the first stage, speech samples were classified as AR or NAR using a parallel configuration of 4 classification channels. Each
channel provided an individual classification result based on a single category of features (G, PS, TEOS or SS). At the second stage the outputs from the 4 channels were combined into the final decision by calculating a weighted sum of the intermediate decisions generated by each channel given in Equation (5.39). The weight values were calculated using Equations (5.36)-(5.38).

Two approaches to the fused classification process were tested: classification with a single set of weights (1W) from both AR and NAR participants and classification with two separate sets of weights (2W), one from the AR participants ($W_{AR}$) and one from the NAR participants ($W_{NAR}$). Table 5.12 shows the resulting weight values averaged over 3 cross validations. Like in the case of a single-channel classification discussed in Section 5.8, both the utterance based (UB) classification and person based (PB) classification were tested for comparison purposes.

The weights values in Table 5.12 reflect contributions of individual feature categories into the formation of the final classification decision based on the weighted score parameter given in Equation (5.39). In the case of the 1W approach, as well as the 2W approach with $W_{AR}$, the order of weights (from the highest to the lowest) was: P, G, S, and TEO. However; for the 2W approach with $W_{NAR}$ the order was different: TEO, G, S, and P. This indicates that the prosodic (P) and glottal (G) features are more highly correlated with the AR speech characteristics, whereas the TEO and glottal (G) parameters show the highest correlation with the NAR characteristics.

Interestingly, in all cases, the glottal parameters appeared high in the list. It is therefore likely that they play very important role in distinguishing between the AR and NAR individuals. This is consistent with similar observations described in [75], [84].
[85], where the glottal features were reported to provide strong enhancement of the accuracy of clinical depression recognition, especially when combined with prosodic or spectral features. Speech emotion recognition results described in [49] also pointed to the high importance of glottal features and their links to psycho-physiological mechanisms that can lead to stress, emotion or depression related changes in the mechanisms of speech phonation.

The effectiveness of weight allocation given in Table 5.12 on the performance of the multi-channel predictive classification system is illustrated in Table 5.13 and Table 5.14, which provide specificity, sensitivity and accuracy values resulting from the utterance based (UB) and the person based (PB) prediction described in Section 5.7.

Table 5.12: Weight Values for the 1W and 2W MCWSC Method

<table>
<thead>
<tr>
<th>Classification Decision Method</th>
<th>Weights Allocation for Individual Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Glottal</td>
</tr>
<tr>
<td>MCWSC</td>
<td></td>
</tr>
<tr>
<td>WAR</td>
<td>0.2435</td>
</tr>
<tr>
<td>WNAR</td>
<td>0.2589</td>
</tr>
<tr>
<td>MCWSC / IW</td>
<td>0.2567</td>
</tr>
</tbody>
</table>

Table 5.13: Depression Prediction Results for the 1W and 2W MCWSC Method Using the Utterance Based (UB) Approach

<table>
<thead>
<tr>
<th>Prediction Method</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCWSC / 2W</td>
<td>65.41</td>
<td>55.84</td>
<td>60.63</td>
</tr>
<tr>
<td>MCWSC / 1W</td>
<td>62.48</td>
<td>45.89</td>
<td>54.19</td>
</tr>
</tbody>
</table>
Table 5.14: Depression Prediction Results for the 1W and 2W MCWSC Method Using the Person Based (PB) Approach

<table>
<thead>
<tr>
<th>Prediction Method</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCWSC / 2W</td>
<td>78.57</td>
<td>66.96</td>
<td>72.77</td>
</tr>
<tr>
<td>MCWSC / 1W</td>
<td>74.70</td>
<td>41.96</td>
<td>58.33</td>
</tr>
</tbody>
</table>

A comparison between the single weight (1W) and the two weight (2W) combined feature classification approach in Table 5.13 and Table 5.14 clearly indicates that the 2W approach provided significantly better performance in terms of the classification accuracy as well as the sensitivity to specificity ratio. This can be explained by the fact that the 2W approach provided different feature ranking orders (see Table 5.12) for the AR weights (W\textsubscript{AR}) and for the NAR weights (W\textsubscript{NAR}), which implies that different types of features are needed to provide classification cues for the AR participants compared with the NAR participants. The 2W approach takes these differences into account by using two different sets of weights and thus reflecting different ranking orders of feature categories for the AR and NAR classes.

Comparing the UB approach of the MCWSC (Table 5.13) with the single-channel classification (SCSC), a significant improvement across all performance evaluation method (sensitivity, specificity and accuracy) was seen when using the 2W method. The 1W approach however provided a rather weak sensitivity to specificity ratio with a low specificity of 46% recorded. The UB approach however did not perform as well as the PB approach which was also noticed in the single-channel classification (SCSC) discussed in Section 5.8.2. Considering the highly challenging nature of speech based prediction of the onset of depression symptoms, the person based (PB) depression
prediction approach with two separate sets of weights (2W) provided very promising results (Table 5.14). The prediction accuracy of almost 73% is exceptionally high for a prediction that depression will develop up to 2 years into the future. It places the system performance well above the pure chance level of 50% for a binary classification. In addition, there is a desirable sensitivity to specificity ratio with a high sensitivity value of almost 79% and smaller specificity value of 67%. This ratio ensures a relatively large true positive numbers (AR classified as AR) at the cost of having higher (but not too high) numbers of false positives (NAR classified as AR). This particular balance between sensitivity and specificity may be not the most desirable for the diagnosis of existing depression, but it can be considered as quite reasonable for the early prediction or screening of people who are likely to develop depression in the near future. Therefore, the results show that the concept of speech based prediction of clinical depression is feasible.

The differences between weights in Table 5.12 are very small indicating that on average all features played equally important roles in the process of depression prediction. However the weight values are used directly in the classification process only with the 1W system. The 2W system makes the decision using the signs of the weighted scores as well as the sign of differences between magnitudes of these scores. The superior performance of the 2W system indicates that the actual differences between weight values may not be as important as the ranking order of features provided by these weights for different classes. It is possible that different features may have performed best with different speech sounds. The weighted multiple-channel system was selecting the best
feature for each speech sound, leading to overall improvement of the classification performance.

In conclusion, the new MCWSC system outperformed the previously tested SCSC (Section 5.8) and MCSC (Section 5.9) methods and led to depression prediction accuracy of up to 73% in the person based (PB) approach.

The following section investigates further improvements to the MCWSC method and introduces a new method of determining the weight values.

5.11 New Optimised Multi-Channel Weighted Speech Classification (OMCWSC) System

This section introduces a new multi-channel classification system using the optimization algorithm (see Section 5.11.3) for an intensive search for an optimized weight allocation for each individual single channel used. Section 5.10 introduced the MCWSC process by determining weight values for each single-channel using a supervised classification, calculating the individual objective function representing a total average squared error between the actual and estimated classes for each output. The normalized values for each objective function are considered as the optimal weights representing individual channels.

This method further improves the weights allocation procedure by an automatic permutated optimization algorithm using the evolutionary search and simulated annealing procedure. This optimization algorithm includes a coarse and fine constrained, weights search which provides a more complete analysis on weights allocation by considering all possible combinations of the weights scale arrangement. The following sub-sections
explains in detail the methodology for the new optimised multi-channel weighted classification system (OMCWSC) and the system performance for depression prediction are discussed.

5.11.1 System Description

Figure 5.12 illustrates the multi-channel classification process in which, M parallel single-channel classifiers provided independent class estimates (prediction results) \( \hat{y}_k(x_i), k = 1, ..., M \) based on \( M \) different types of features. These intermediate single-channel estimates were then combined into a weighted sum \( r(x_i) \) given as

\[
    r(x_i) = \sum_{k=1}^{M} W_{k}^{opt} \hat{y}_k(x_i)
\]

(5.40)

Where \( W_{k}^{opt} \) denote optimised weight values. The process used to determine these values is described in Section 5.11.2. Using the \( r(x_i) \) values, the final prediction results in a form of the \( \hat{y}(x_i) \) values (equal to +1 for AR and -1 for NAR) were determined using a decision matrix look up.

When searching for suitable weights’ values, two different cases were investigated; one with a single set of weights (1W) estimated from data representing both AR and NAR classes and one with two separate sets of weights (2W) with one set estimated using only the AR data and one - using only the NAR data. In each case, a separate decision matrix was designed. Table 5.15 shows the decision matrix used in the case of a single set of weights. The procedure was in this case simply assigning the AR
class to the positive values of \( r(x_i) \) and the NAR to the negative values of \( r(x_i) \). In the case of two separate sets of weights (2W), the decision process was based on the set of rules illustrated in Table 5.16. For each of the test samples \( x_i \), the classification procedure was in this case generating two different values of \( r(x_i) \): \( r_{AR}(x_i) \) and \( r_{NAR}(x_i) \). Where the first one was calculated using a set of weights representing exclusively the AR class and the second - using a set of weights representing exclusively the NAR class. The final decision in the 2W case depended on the signs and relative values of \( r_{AR}(x_i) \) and \( r_{NAR}(x_i) \). When both parameters had the same sign, the final decision was automatically assigning AR for positive values and NAR for negative values.

Providing that each single channel worked with features that were previously selected to provide high correlation with the risk for depression, the choice of the weight values played a key role in achieving high multi-channel prediction accuracy.

Table 5.15: OMCWSC DECISION MATRIX FOR A SINGLE SET OF WEIGHTS (1W)

<table>
<thead>
<tr>
<th>sign of ( r(x_i) )</th>
<th>Value of ( \hat{g}(x_i) )</th>
<th>Predicted Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+1</td>
<td>AR</td>
</tr>
<tr>
<td>-</td>
<td>-1</td>
<td>NAR</td>
</tr>
</tbody>
</table>

Table 5.16: OMCWSC DECISION MATRIX FOR A TWO SET OF WEIGHTS (2W)

<table>
<thead>
<tr>
<th>sign of ( r_{AR}(x_i) )</th>
<th>sign of ( r_{NAR}(x_i) )</th>
<th>If condition</th>
<th>Value of ( \hat{g}(x_i) )</th>
<th>Final prediction result</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>Undetermined</td>
<td>+1</td>
<td>AR</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>(</td>
<td>r_{AR}(\tilde{x_i})</td>
<td>&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(</td>
<td>r_{AR}(\tilde{x_i})</td>
<td>&lt;</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>(</td>
<td>r_{AR}(\tilde{x_i})</td>
<td>&lt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(</td>
<td>r_{AR}(\tilde{x_i})</td>
<td>&gt;</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>Undetermined</td>
<td>-1</td>
<td>NAR</td>
</tr>
</tbody>
</table>
5.11.2 Finding the optimal weight values

The weight coefficients were determined using a combined classification and weight optimization procedure illustrated in Figure 5.13. The classification process performed by individual channels had a supervised character which means that, for each data sample \(x_i\) \((i=1,\ldots,N)\), the actual class \(y(x_i)\) to which this sample belonged was known. This knowledge was needed to assess the system performance for a given set of weight values and thus to find a set of weight values which provided the best performance.

Given a speech sample \(x_i\), a set of weight values \(W = \{W_1, \ldots, W_M\}\) and classification outputs from each channel \(\hat{y}_k(x_i), k=1,\ldots,M\), the classification output \(g(x_i)\) from the multi-channel system was determined as

\[
g(x_i) = \sum_{k=1}^{M} W_k \hat{y}_k(x_i) \tag{5.41}
\]

Using an iterative global optimization procedure, an extensive two-stage search was performed to determine a set of weight values \(W^{opt} = \{W_1^{opt}, \ldots, W_M^{opt}\}\) that minimized the objective function \(f_{obj}(W)\) defined as

\[
f_{obj}(W) = \sum_{i=1}^{N} [g(x_i) - y(x_i)] \tag{5.42}
\]

This search was initialised with weight values set to

\[
W_k^{init} = \frac{1}{M}, \quad \text{where}, k = 1, \ldots, M \tag{5.43}
\]
The data used to determine the weight values was separate from the sets used to train and the set used to test the system. The two-stage weight optimization procedure consisted of

a) Coarse, constrained search, and

b) Fine, constrained search

\[ r(x_i) = \sum_{k=1}^{M} W_{k}^{opt} \hat{y}_k(x_i) \]

\[ \hat{g}(x_i) \rightarrow \text{Decision Matrix} \]

**Figure 5.12:** Optimized multi-channel weighted speech classification (OMCWSC) system testing stage.
5.11.2.1 Coarse, Constrained Search

Meta-heuristic global optimization procedures provide very powerful tool for searching large multivariate spaces, however when applied in an unconstrained way, the time needed to converge to desirable solutions is in most cases very long [11]. In order to improve the search efficiency and thus to reduce the time needed to determine a “good” set of weights, the initial search was conducted within constrains imposing specific rank of the M classification channels. Given the number of M channels, a permutation matrix was generated with each row representing (in the descending order) the channels’ rank.

**Figure 5.13:** Training process for finding the optimal weight vector in the OMCWSC approach.
determined by the corresponding weight values. Table 5.17 shows an example of the permutation matrix for three channels \((M=3)\). The optimization process was then run separately within each of the channel ranking constrains represented by the matrix rows to minimise the objective function in Equation (5.42). The solutions corresponding to each row of the permutation matrix were used to determine the order of weight values corresponding to the top ranking row. This order was used to constrain the next step of the optimization process.

### 5.11.2.2 Fine, Constrained Search

After coarse search, the best solution was used as an initial point to perform a fine-tuning of the weight values. However, the optimization process was in this case constrained to keep the channel rank of weight values as determined by the best performing row of the permutation array used during the coarse search. In addition, the search was conducted in a more intensive way by adjusting the algorithm’s parameters, controlling the algorithm sensitivity and time used to explore the parameter’s space. The optimised vector of weights produced by the fine search was then applied in the multi-channel classification procedure as illustrated in Figure 5.12.

#### Table 5.17: An Example of Permutation Matrix for 3 Channels \((M=3)\)

<table>
<thead>
<tr>
<th>Runs</th>
<th>Channel orientations for weight allocation</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Highest</td>
<td>Second Highest</td>
<td>Lowest</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
5.11.3 Optimization algorithm

The optimization (or function minimization) algorithm [Mitchell, 2005] used to determine the optimised weights’ values was a hybrid of two meta-heuristic strategies, evolutionary search [11], [45], [146] and simulated annealing (SA) [64], [146]. In addition to the Boltzmann acceptance criterion typically used in SA for the monitoring of the objective function value (energy) [64], the algorithm restricted the search step size in the vector space to a limited range decreasing exponentially with the number of iterations. This combined energy/range control process resulted in an improvement of the search efficiency and better detection of the local “downhill movement” possibilities. In addition, the evolutionary aspect of the algorithm allowed generating large populations of possible solutions and keeping only the “best” solutions to produce offspring populations during the subsequent iterations. At the beginning, the algorithm was scanning a wide range of the weight values and was allowed to accept solutions leading to decrease or to a small increase of the objective function value. As the algorithm progressed, the probability of accepting solutions leading to higher values of the objective function was gradually decreasing to zero. At the same time, the search range was becoming increasingly confined to a small space around the final solution thus preventing the algorithm from redundant and time consuming explorations of wide range solutions.

5.11.4 Testing the OMCWSC system

Test results presented in Section 5.10 have shown that the multi-channel weighted system (MCWSC) using a number of single-channel classifiers, with each channel processing
only one type of features provided significant improvement over the individual performances of these channels (shown in Section 5.8).

The normalised weight allocation to individual channels used in the MCWSC method was based on an arbitrary distribution of values proportional to the individual channels’ performances. The OMCWSC system improved the weight allocation procedure and used a global optimization method to derive the weight values which minimize the classification error expressed in Equation (5.42).

The OMCWSC performance was tested using the following multi-channel configurations:

1) 5 channels representing TEOG-SB1, TEOG-SB2, TEOG-SB3, TEOG-SB4 and TEOG-SB5
2) 4 channels representing TEOG-SB2, TEOG-SB3, TEOG-SB4 and TEOG-SB5
3) 3 channels representing TEOG-SB3, TEOG-SB4 and TEOG-SB5
4) 3 channels representing G, PS and TEOG-SB3+SB4+SB5

Cases 1-3 investigated the performance of the system performance with TEOG features representing different frequency bands, whereas the fourth case (G, PS and TEOG-SB3+SB4+SB5) combined the best performing single channel features (see Section 5.8.2) with the best performing TEOG sub-bands (see Section 5.8.4).

All configurations were tested within the person based (PB) scheme using two approaches:

1) Classification with a single set of weights (1W) using the decision matrix in Table 5.15;
2) Classification with two separate sets of weights (2W), one from the AR participants (WAR) and one from the NAR participants (WNAR) using the decision matrix in Table 5.16.

5.11.4.1 Testing the OMCWSC with TEOG features within different sub-bands

Table 5.18, Table 5.19 and Table 5.20 show the classification results for TEOG features representing different frequency bands.

**Table 5.18: Depression Prediction Results for the OMCWSC System with TEOG-SB1-SB5 (100Hz – 5500Hz) Features**

<table>
<thead>
<tr>
<th>Prediction Method</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMCWSC/2W</td>
<td>55.36</td>
<td>61.91</td>
<td>58.63</td>
</tr>
<tr>
<td>OMCWSC/1W</td>
<td>55.36</td>
<td>57.14</td>
<td>56.25</td>
</tr>
</tbody>
</table>

**Table 5.19: Depression Prediction Results for the OMCWSC System with TEOG-SB2-SB5 (700Hz - 5500Hz) Features**

<table>
<thead>
<tr>
<th>Prediction Method</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMCWSC/2W</td>
<td>64.29</td>
<td>61.31</td>
<td>62.78</td>
</tr>
<tr>
<td>OMCWSC/1W</td>
<td>63.69</td>
<td>48.80</td>
<td>56.25</td>
</tr>
</tbody>
</table>

**Table 5.20: Depression Prediction Results for the OMCWSC System with TEOG-SB3-SB5 (1300Hz – 5500Hz) Features**

<table>
<thead>
<tr>
<th>Prediction Method</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMCWSC/2W</td>
<td>68.45</td>
<td>70.24</td>
<td>69.35</td>
</tr>
<tr>
<td>OMCWSC/1W</td>
<td>63.69</td>
<td>61.91</td>
<td>62.80</td>
</tr>
</tbody>
</table>
The objective was to further investigate the previous finding described in Section 5.8.4 where, the single classification channels with TEOG features calculated within higher frequency bands provided better classification compared to the TEOG features derived from the lower frequency bands.

The following general observations can be made based on the results in Table 5.18-Table 5.20:

1) The higher TEOG feature provides better performance of the OMCWSC system when the low frequency components are removed.

2) The highest performance of the OMCWSC system was achieved when using the TEOG features calculated within 1300Hz-5500Hz. This applies for both 1W (63%) and 2W (69% accuracy) approach.

3) The results for 2W approach were in all cases higher than for the 1W approach.

This is consistent with the MCWSC results in Section 5.10.

The highest classification accuracy of 69% was achieved in for 3-channel system with TEOG features calculated within 1300Hz-2400Hz, 2400Hz-3500Hz and 3500Hz-5500Hz implemented with the 2W approach. The sensitivity to specificity ratio was 68/70 placing the results at quite desirable position.

It was noticed that in all cases, the 2W approach clearly outperformed the 1W approach. This can be due to the fact that the 2W approach provided different orders of weight allocation for the AR weights (W_{AR}) and for the NAR weights (W_{NAR}), which implies that different types of features could be more important in providing classification cues for the AR participants and for the NAR participants.
In addition, the 2W system makes the decision using the signs of weighted scores as well as the sign of differences between magnitudes of these scores (see Table 5.16), whereas weighted score values in the 1W system are used directly in the classification process without any comparison (see Table 5.15).

5.11.4.2 Testing the OMCWSC with G, PS and TEOG (SB3-SB5) features

This section investigates the performance of the OMCWSC structure with three channels (G, PS and TEOG (SB3-SB5)) combining the best performing single channel features (see Section 5.8.2) with the best performing TEOG sub-bands (see Section 5.8.4). The results are presented in Table 5.21.

Findings in Section 5.8 suggested that the glottal (G) and prosodic (PS) parameters were the best performing single-channel features. The G features on their own provided classification accuracy of 69% and sensitivity to specificity ratio of 76%/63%. The PS features came second with a slightly lower accuracy of 63% and the sensitivity to specificity ratio of 73%/53%. The TEOG (SB3-SB5) features on the other hand led to a very similar performance to 68% accuracy and sensitivity to specificity ratio of 66/69.

Table 5.21 shows that these three best performing single channels when combined into the OMCWSC system produced very desirable results. Although when compared to results in Table 5.14 (MCWSC system) achieves only a slight improvement, it was noticeable that the sensitivity to specificity ratio was more balanced. Comparing the PB, 2W approach for both systems, the difference in sensitivity to specificity ratio for MCWSC was 11% compared to OMCWSC with a difference of only 7%. More evidently in the PB, 1W approach, the difference for MCWSC was 32% compared to 19% for the
OMCWSC approach. The optimization did not only achieve better accuracy improvement but also a less skewed ratio performance.

The depression prediction accuracy of almost 74% (for the PB, 2W approach) is exceptionally high providing that the first symptoms of depression detectable by diagnostic questionnaires appeared 2.5 years after the tested speech recordings were made. It places the system performance 24% above the pure chance level of 50% for a binary classification.

In addition, there is a desirable sensitivity to specificity ratio with both high sensitivity value of almost 77% and specificity value of 70%. This ratio ensures a relatively large true positive numbers (AR classified as AR) at the cost of having higher (but not too high) numbers of false positives (NAR classified as AR). This way a reasonable balance between sensitivity and specificity was achieved.

These results further confirm that the concept of a speech based prediction of clinical depression is feasible. Considering the highly challenging nature of speech based prediction of the onset of depression symptoms, the multi-channel optimized weighted speech classification OMCWSC system with two separate sets of weights (2W) provided very promising results.

<table>
<thead>
<tr>
<th>Prediction Method</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMCWSC/2W</td>
<td>76.79</td>
<td>70.24</td>
<td>73.51</td>
</tr>
<tr>
<td>OMCWSC/1W</td>
<td>76.79</td>
<td>57.74</td>
<td>67.26</td>
</tr>
</tbody>
</table>
5.12 Summary and Conclusions

This chapter examined the usefulness of acoustic speech parameters in the identification of adolescents at high imminent risk of clinical depression. In the first stage, the study investigated single-channel speech classification approaches based on a single type of acoustic features. In the second stage, new multi-channel classification systems which combine classification outcomes of single-channel components into the final classification decision were proposed. The proposed methods were tested with conversational speech data collected from 30 adolescents diagnosed as non-depressed at the time when the speech data was acquired. Two years later the follow up diagnosis confirmed that 15 out of the 30 individuals had developed symptoms of clinical depression.

It was demonstrated that acoustic speech parameters can be used to predict the risk for depression in adolescents 2.5 years before the diagnostic criteria for the full-blown disorder can be used to determine early signs of Major Depression in adolescents. The prediction accuracy achieved in the best case was 74%, with the specificity to sensitivity ratio of 70%/77%.

Other, more detailed outcomes of the investigation can be summarised as follows:

1) It was observed that the single-channel classification was effective in predicting depression with a desirable specificity to sensitivity ratio and accuracy higher than chance level only when using glottal (G), (PS) or TEOG (within 1300Hz-5500Hz) features.

2) In all classification systems the person based (PB) approach provided better performance than the utterance based (UB) approach.
3) The proposed multi-channel methods outperformed the single-channel approaches.

4) The best prediction result of 74% accuracy, 77% sensitivity and 70% specificity was achieved with the optimised multichannel weighted speech classification (OMCWSC) method comprised of 3 single-channels processing the G, PS and TEOG (SB3-5) features.
Chapter Six:

SUMMARY AND FUTURE WORK

6.1 Research Summary
The study investigated prediction of Major Depression in adolescents using two approaches; classification of facial images and classification and acoustic speech parameters.

The prediction task was facilitated by the nature of available audiovisual recordings made when all participants were professionally diagnosed as normal healthy adolescents with no current or previous episodes of depression. These recordings were accompanied by clinical diagnosis made 2.5 years after the recordings were taken; this diagnosis showed which participants developed symptoms of Major Depression. Using this information it was possible to classify the audiovisual recordings into two categories: participants “At Risk” (AR) for depression and participants “Not At Risk” (NAR) for depression. The study was based on 15 participants (6 male and 9 female) representing the AR group and 15 participants (6 male and 9 female) representing the NAR group. To our knowledge, this was the first study of its kind.

In order to determine if facial image features can distinguish between the AR and NAR classes, two different types of features: eigenfaces and fisherfaces were
investigated. In both cases the Nearest Neighbour (NN) classifier was used to generate the prediction outcome. The results were not satisfactory. The highest achieved prediction accuracy was 61% (when using the fisherfaces, PSI session and person dependent approach) with sensitivity to specificity ratio of 65%/58%. Although this result was above the pure guessing level of 50%, it was too low to provide a definite answer to the question of whether it is possible to predict depression from facial images. Further investigations into different methods of facial features extraction and possibly more advanced multi-channel classification approaches are needed to give a more definite answer.

These conclusions led to the investigation of vocal features from adolescent’s voice recordings as possibly better predictors of depression than the facial features.

The study began with an investigation of classical, single-channel approaches, where the modelling process and the classification decision procedures were based on a single type of features (Glottal, Prosodic, TEO or Spectral) and proceeded into new, advanced multi-channel systems with each channel performing independent classification based on a different type of features.

The following primarily research questions were setup:

1) Is it possible to predict the risk for depression from acoustic speech parameters 2.5 years before the full blown symptoms occur?

2) What kind of acoustic parameters provide the best discrimination between the “At Risk” (AR) and “Not AT RISK” (NAR) adolescents?

3) What is the most efficient speech modeling and classification procedure for an early prediction of depression?
The results provided the following answers:

1) Yes, it is definitely possible to predict the risk for depression from acoustic speech parameters 2.5 years before the full blown symptoms occur. The experiments showed that the speech based classification accuracy can reach 74%.

2) Statistical tests on a wide range of glottal, prosodic, TEO and spectral parameters have shown that in all cases except jitter (from the prosodic category), there was a statistically significant difference between mean values for the AR and NAR groups. The efficiency of individual types of features was tested using a classical single-channel classification approach. It was found that out of the four categories (glottal, prosodic, TEO and spectral), the glottal features were the most efficient in discriminating between AR and NAR individuals. The best prediction accuracy provided by the glottal features was 69% with good sensitivity to specificity ratio of 76%/62%.

3) The study proposed a new multi-channel weighted speech classification (MCWSC) method. It was found that the optimised version of this method, called the optimised multi-channel weighted speech classification (OMCWSC) system was the most efficient in prediction of risk for depression. The OMCWSC system comprises of a number of single-channels making individual predictions based on a single type of features. The weighted outcomes from these individual channels are combined into the final prediction result. When implemented in the person based (PB), two weights (2W) approach with 3 single-channels processing glottal (G), prosodic (PS) and TEOG (within 1300Hz-5500Hz) features, the OMCWSC achieved 74% prediction accuracy, 77% sensitivity and 70% specificity.
6.2 Future Work

The thesis presented first of its kind research of image and speech based early prediction of major depression.

The study was facilitated but also limited by the nature of the available data base of audiovisual recordings. Therefore the conclusions need to be validated and confirmed by future similar studies.

It would be highly desirable for the future studies to take into consideration the following issues:

1) There is a need for a large database of audiovisual recordings made before the depression symptoms occur and also a few years later when the symptoms are already present. The limitation of the data base used in this study was the recordings were made only before the depression symptoms occurred. It would be very beneficial to compare the image and speech characteristics before and after symptoms of depression occur.

2) The available size of the testing sample of 15 AR and 15 NAR participants was relatively very small and much larger samples should be tested.

3) Since the adolescent’s speech differs from the speech of adults, investigations into different age groups are needed.

4) The study was limited to one type of the depressive disorder known as the Major Depression. Investigations into the whole spectrum of different types of depression disorders are needed.
5) Facial expressions, as well as the speech acoustic characteristics vary across languages and cultures. Future investigations of these variations in relation to prediction of depression are needed.

6) A combined approach merging the facial and speech acoustic information into the multi-channel depression prediction system should be also considered in future studies.
## APPENDIX A

**Table A1:** Representation of the codes used in the LIFE coding system

<table>
<thead>
<tr>
<th>Content codes</th>
<th>Expression</th>
<th>Affect codes</th>
<th>Emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Validation</td>
<td>0</td>
<td>Contempt</td>
</tr>
<tr>
<td>14</td>
<td>Affection</td>
<td>1</td>
<td>Anger</td>
</tr>
<tr>
<td>15</td>
<td>Solicitous</td>
<td>2</td>
<td>Anxious</td>
</tr>
<tr>
<td>16</td>
<td>Humor</td>
<td>3</td>
<td>Dysphoric</td>
</tr>
<tr>
<td>18</td>
<td>Approve</td>
<td>4</td>
<td>Pleasant</td>
</tr>
<tr>
<td>21</td>
<td>Complaint</td>
<td>5</td>
<td>Neutral</td>
</tr>
<tr>
<td>23</td>
<td>Cruelty</td>
<td>6</td>
<td>Happy</td>
</tr>
<tr>
<td>24</td>
<td>Negative substance</td>
<td>7</td>
<td>Caring</td>
</tr>
<tr>
<td>25</td>
<td>Provoke</td>
<td>8</td>
<td>Whine</td>
</tr>
<tr>
<td>26</td>
<td>Annoy-disrupt</td>
<td>9</td>
<td>Belligerence</td>
</tr>
<tr>
<td>28</td>
<td>Disagree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Command</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Command unaccountable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Comply</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Noncomply</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>Self-statement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>Self-positive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>Self-complaint</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>Problem statement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>Propose solution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>Teach</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>Guidance tactic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>Attend</td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>Talk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>Inaudible</td>
<td></td>
<td></td>
</tr>
<tr>
<td>98</td>
<td>Dummy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
REFERENCES


